UNIVERSIDADE FEDERAL DO PARANÁ

JULIANE REZENDE MERCER

ESTRUTURA GENÉTICA DE UMA POPULAÇÃO SELECIONADA DE *Pinus taeda* Linnaeus

CURITIBA

2011

JULIANE REZENDE MERCER

ESTRUTURA GENÉTICA DE UMA POPULAÇÃO SELECIONADA DE *Pinus taeda* Linnaeus

Tese apresentada ao Programa de Pós-Graduação em Genética, Universidade Federal do Paraná, como requisito à obtenção do título de Doutor em Ciências Biológicas, Área de concentração: Genética.

Orientadora: Prof.ª Dra. Marina Isabel Mateus de Almeida
Co-Orientador: Prof. Dr. Antonio Rioyei Higa
Co-Orientadora: Dra. Milena de Luna Alves Lima
Co-Orientadora: Dra. Juliana Vitória Messias Bittencourt

CURITIBA

2011

Ministério da Educação
UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE CIÊNCIAS BIOLÓGICAS
DEPARTAMENTO DE GENÉTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA

Ata da defesa de tese de doutorado de JULIANE REZENDE MERCER

Aos seis dias do mês de abril do ano de dois mil e onze, foi realizada na sala 2 do Departamento de Botânica, da Universidade Federal do Paraná, a defesa de tese da doutoranda **JULIANE REZENDE MERCER** intitulada: "**Estrutura Genética de uma População Selecionada de *Pinus Taeda* L.**". A abertura teve início às quatorze horas pelo Professor Doutor **Ricardo Lehtonen Rodrigues de Souza**, Vice-Coordenador do Programa de Pós-Graduação em Genética, que passou a palavra à Professora Doutora **Marina Isabel Mateus de Almeida**, do Departamento de Genética da UFPR, Orientadora *da candidata e Presidente da banca examinadora, a qual apresentou ao público presente os* membros da banca examinadora e passou a palavra à aluna, para que fizesse a apresentação da sua tese. Após a apresentação oral, a Professora Doutora **Marina Isabel Mateus de Almeida** passou a palavra ao primeiro examinador Professor Doutor **Edson Perez Guerra**, da PUC/PR. Em seguida, passou a palavra à segunda examinadora, Doutora **Ananda Virgínia de Aguiar**, da EMBRAPA Florestas. Na sequência, o terceiro examinador, Professor Doutor **Antonio Rioyei Higa**, do Departamento de Ciências Florestais da UFPR, arguiu a candidata. Em seguida, a quarta examinadora, Professora Doutora **Milena de Luna Alves Lima**, do Departamento de Fitossanidade da UFPR, fez suas considerações. Na sequência, a quinta examinadora, Professora Doutora **Chirlei Glienke**, do Departamento de Genética da UFPR, fez suas considerações. Findas as *arguições pelos demais membros da banca, a Professora Doutora* **Marina Isabel Mateus de Almeida** fez uma rápida apreciação das conclusões mais importantes dos debates realizados e comunicou que a banca examinadora iria proceder à discussão para atribuição dos conceitos, reunindo-se em sessão secreta. Os trabalhos foram interrompidos por cinco minutos. Após, foram proclamados os conceitos atribuídos pela banca examinadores, a seguir descritos: Professor Doutor **Edson Perez Guerra**, conceito "B"; Doutora **Ananda Virgínia de Aguiar**, conceito "B"; Professor Doutor **Antonio Rioyei Higa**, conceito "B"; Professora Doutora **Milena de Luna Alves Lima**, conceito "B"; Professora Doutora **Chirlei Glienke**, conceito "B"; Professora Doutora **Marina Isabel Mateus de Almeida**, conceito "B"; com o conceito médio final "B". Tendo cumprido o que dita o artigo trinta e cinco das Normas Internas do Programa, o candidato cumpriu os requisitos para obtenção do grau de Doutor em Ciências Biológicas, área de concentração Genética. Como não havia nada mais a ser tratado, o Professor Doutor **Ricardo Lehtonen Rodrigues de Souza**, após informar ao candidato que ele tem, a partir desta data, até trinta dias para a entrega da versão definitiva de sua tese e comprovante de submissão de pelo menos um artigo científico para um periódico de circulação internacional com cópia do(s) artigo(s), deu por encerrada a sessão. Eu, **Ricardo Lehtonen Rodrigues de Souza**, Vice-Coordenador do Programa de Pós-Graduação em Genética, lavrei a presente ata, a qual assino juntamente com os senhores examinadores. Curitiba, seis de abril de dois mil e onze.

Centro Politécnico - Jardim das Américas - Caixa Postal 19071 - CEP 81531-980 - Curitiba, Brasil.
Tel (+41) 33611587 / 33611684 – *Fax* (+41) 33611793 – *e-mail* ppg-gen@ufpr.br

Ministério da Educação
UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE CIÊNCIAS BIOLÓGICAS
DEPARTAMENTO DE GENÉTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA

PARECER

Os abaixo-assinados, membros da Banca Examinadora da defesa de tese de Doutorado a qual se submeteu **JULIANE REZENDE MERCER,** para obtenção do título de Doutora em Ciências Biológicas na área de Genética pela Universidade Federal do Paraná, no Programa de Pós-Graduação em Genética, são de parecer que se confira ao candidato o conceito "B".
Secretaria da Coordenação do Programa de Pós-Graduação em Genética do Setor de Ciências Biológicas da Universidade Federal do Paraná.

Curitiba, 6 de abril de 2011

Professor Doutor Edson Perez Guerra
PUC/PR – Membro Titular

Doutora Ananda Virgínia de Aguiar
EMBRAPA/Florestas – Membro Titular

Professor Doutor Antonio Rioyei Higa
Depto Ciências Florestais/UFPR – Membro Titular

Professora Doutora Milena de Luna Alves Lima
Depto Fitossanidade/UFPR – Membro Titular

Professora Doutora Chirlei Glienke
Depto GenéticaUFPR – Membro Titular

Professora Doutora Marina Isabel Mateus de Almeida
UFPR – Orientadora e Presidente da banca

Visto

Professor Doutor Ricardo Lehtonen Rodrigues de Souza
Vice-Coordenador do Programa de Pós Graduação em Genética

Centro Politécnico - Jardim das Américas - Caixa Postal 19071 - CEP 81531-980 - Curitiba, Brasil.
Tel (+41) 33611587 / 33611684 – *Fax* (+41) 33611793 – *e-mail* ppg-gen@ufpr.br

SUMÁRIO

LISTA DE FIGURAS E TABELAS

GLOSSÁRIO

BP: *Breeding Population* ou População Selecionada (An.I).

C=*x*: Número (*x*) de grupos genéticos, ou *clusters*, ou fragmentos populacionais, determinados na análise bayesiana assumindo endogamia.

c'$_x$: Número do grupo genético fixado em análises bayesianas consecutivas determinadas assumindo endogamia. Sendo o índice *x*: 1, 2, ..., *i* (*i*: é igual soma total de grupos genéticos).

Δ*K*: é a ordem de significância atribuída para o parâmetro analisado estimada de acordo com a estatística *ad hoc* de EVANNO.

ΔK_m: é a ordem de significância atribuída para o parâmetro analisado estimada de acordo com uma modificação realizada neste estudo, que foi efetuada na equação da estatística *ad hoc* de EVANNO, que substituiu os valores de médias da distribuição de probabilidades *a posteriori* do parâmetro K, para medianas.

Estatística *ad hoc*: É uma estatística complementar para determinar a probabilidade *a posteriori* mais provável dentre uma distribuição de probabilidades *a posteriori* resultantes para um determinado parâmetro.

f: Coancestria de alelos dentro de indivíduos em relação à subpopulação em que ele ocorre, o equivalente a F_{IS}

F: Coancestria de alelos dentro de indivíduos em relação a toda a população, o equivalente à $F_{IT.}$

F_{IS}*:* A correlação dos alelos dentro de indivíduos em relação à subpopulação em que ele ocorre; equivalente ao desvio da expectativa de equilíbrio de Hardy-Weinberg para a média da freqüência genotípica dentro das populações.

F_{ST}*:* Correlação entre alelos escolhidos aleatoriamente dentro da subpopulação em relação a toda a população; equivalente à proporção da diversidade genética devido às diferenças nas freqüências alélicas entre as populações.

F_{IT}*:* Correlação dos alelos dentro de indivíduos em relação à população total, equivalente ao desvio das expectativas de Hardy-Weinberg das freqüências genotípicas em relação a toda a população.

ϕ_{IS} ("*fi*"): O excesso de similaridade de alelos dentro de indivíduos em relação à subpopulação em que ele ocorre; análogo à F_{IS}

ϕ_{ST} ("*fi*"): O excesso de similaridade entre os alelos escolhidos aleatoriamente dentro da mesma subpopulação em relação à população total, ou o equivalente à proporção da diversidade genética atribuídas às diferenças entre as populações; análogo ao F_{ST}. ϕ_{ST} : obtidos de AMOVA é usado para dados haplotípicos, e exige uma medida da distância evolutiva entre todos os pares de haplótipos.

ϕ_{IT} ("*fi*"): O excesso de similaridade de alelos de um indivíduo em relação a toda a população; análogo à F_{IT}

FP: Floresta Plantada.

GDA: *Genetic Data Analysis* software, de Kent E. Holsinger and Paul O. Lewis, que estima Estatistica-F para marcadores dominantes e codominantes.

HW: Equilíbrio de Hardy-Weinberg.

K: Grupos genéticos, ou cluster, ou fragmento populacional.

K=*x*: Número (*x*) de grupos genéticos, ou *clusters*, ou fragmentos populacionais, determinados na análise bayesiana assumindo equilíbrio de HW.

K_x: Número do grupo genético fixado em análises bayesianas consecutivas determinadas assumindo equilíbrio de HW. Sendo o índice *x*: 1, 2, ..., *i* (*i*: é igual ao número total de grupos genéticos definidos para a população).

k'_x: indica que K_x está sendo decomposto em níveis inferiores de sub-grupos genéticos.

LD: *Linkage Disequilibrium* ou Desequilíbrio de Ligação.

MK=*x*: Número (*x*) de grupos genéticos, ou *clusters*, ou fragmentos populacionais, determinados na análise bayesiana assumindo equilíbrio de HW para a geração materna.

MK_x: Número do grupo genético fixado em análises bayesianas consecutivas determinadas assumindo equilíbrio de HW para a geração materna. Sendo o índice *x*: 1, 2, ..., *i* (*i*: é igual ao número total de grupos genéticos definidos para a população do progenitor feminino).

Open-sib ou *Open-pollinated progeny*: é a denominação dada a uma progênie formada por polinização aberta. Em espécies de plantas alógamas monoicas, o órgão reprodutor feminino pode ser fertilizado por fontes distintas de pólen, de modo que, tem-se o controle somente do lado materno. *Full-sib progeny* ou progênie de polinização controlada ocorre quando o órgão reprodutor feminino é fertilizado com uma única (ou conhecida) fonte de pólen, e este é protegido para que outros pólen não o possam também fertilizar; nesse caso, têm-se o controle tanto materno quanto paterno.

Origem: È o local geográfico onde a espécie ocorre naturalmente.

OS: *Open-sib* ou progênie de polinização aberta.

PCR: *Polymerase Chain Reaction* ou Reação em Cadeia de Polimerase.

$[Pr(K \mid X)]$: Distribuição de probabilidade *a priori* do parâmetro K.

$[Pr(X \mid X)]$: Distribuição de probabilidade *a posteriori* do parâmetro K.

QTL: *Quantitative Trace Locus* ou *Locus* de Carácter Quantitativo.

Procedência da semente: é o local geográfico onde a semente foi coletada, podendo ou não, ser na origem.

R_{ST} : O excesso de similaridade para os alelos selecionados aleatoriamente dentro da subpopulação em relação a toda a população, análogo ao F_{ST}. R_{ST} estimado para microssatélites, e exige que os alelos sejam rotulados de acordo com o número de unidades repetidas contidas no *locus*.

θ ("*thêta*"): Coancestria de alelos escolhidos ao acaso dentro da mesma subpopulação em relação a toda a população, o equivalente a $F_{ST.}$

SAM: *Marker Assisted Selection* ou Seleção Assistida por Marcadores.

SSR: *Simple Sequence Repeats*, sequencias microssatélites.

SNP: *Single Nucleotide Polymorphism*.

AGRADECIMENTOS

Agradeço primeiramente a Deus e ao apoio dos meus pais: Mercer, Diva e Vera. Agradeço a compreensão, principalmente nos momentos difíceis deste trabalho e também ao apoio do meu marido Cristiano e dos meus filhos João Rafael, Matheus e Francisco.

Agradeço às Instituições de fomento à pesquisa: FINEP e Fundação Araucária que tornaram possível a execução desta pesquisa.

Agradeço ao Laboratório de Melhoramento Florestal do Departamento de Ciências Florestais DECIF, particularmente pela confiança em mim depositada para a realização desta pesquisa pelo coordenador do LAMEF Dr. Antonio Rioyei Higa, e pelas coordenadoras das metafísicas Dra. Juliana Vitória Messias Bittencourt e Dra. Milena de Luna Alves Lima.

Agradeço ao incentivo e aos ensinamentos que me foram transmitidos pela minha orientadora, a professora Marina Isabel Mateus de Almeida e Co-orientadora Dra. Milena de Luna Alves Lima.

Agradeço aos professores e ao Programa de Pós-Graduação de Genética PPG-GEN, por todo suporte que me foi dedicado, em especial aos professores Juarez Gabardo, Chirlei Glienke, Lygia Terasawa.

Agradeço o companheirismo dos colegas de doutorado: Paula Rachel Corrêa, Juliana Zanetti, Carolina Andrade, Josiane Guimarães e todos os outros que estiveram muito presentes em várias fases do curso.

Agradeço ao grande empenho das estagiárias do Lamef – estudantes de Engenharia Florestal, que auxiliaram na fase de extração do DNA genômico: Nicole Munhoz, Adriane Partala e Thaís Vagaes. Agradeço também todo o auxílio e suporte na agilidade de aquisição dos reagentes que foram imprescindíveis ao bom caminhamento de todas as atividades de bancada por parte da secretaria do LAMEF, representada por Dona Carmem Ceccon.

RESUMO

Espécies florestais de rápido crescimento suprem de forma sustentável a demanda por biomassa lenhosa, reduzindo a pressão sobre as florestas naturais. No entanto, reflorestamentos com *Pinus* vêm produzindo madeira juvenil de baixa qualidade. Assim, o melhoramento genético para características relacionadas à qualidade da madeira ou outras de difícil mensuração, torna-se prioritário aos programa de melhoramento florestal. Fortes expectativas têm sido direcionadas aos estudos de genética de associação, via genes candidatos, para estimar a correlação genética e fenotípica entre os alelos polimórficos e as características de interesse. Tal estratégia representaria uma possibilidade de predizer ganhos a partir da implantação de um programa baseado em seleção assistida por marcadores. Contudo, a existência de estrutura genética espúria é apontada na literatura como uma das maiores causas de viés estatístico nos estudos de mapeamento de associação. Um dos maiores desafios, mesmo com o avanço de todas as tecnologias genômicas e analíticas, é a definição de quais são os verdadeiros grupos genéticos existentes dentro de uma determinada população. Definir os grupos genéticos em uma população artificial torna-se ainda mais desafiador, uma vez que não se sabe ao certo o histórico da origem dos indivíduos que compõem essas populações. A presente pesquisa parte da hipótese que, definindo o número de K (ou de grupos genéticos) mais provável para representar a população, e, apontando quais indivíduos da população fazem parte de cada grupo, seja a estratégia mais eficiente para conter o viés estatístico em estudo de mapeamento de associação. Partindo-se dessa hipótese, o objetivo foi caracterizar a estrutura genética de uma população selecionada de *Pinus taeda* em busca do K mais provável, com o propósito de prepará-la para um estudo de mapeamento de associação. A população foi formada por 1130 indivíduos com 12 anos de idade, que representam 120 famílias de polinização aberta, e que haviam sido selecionados para melhor desempenho em volume de madeira produzida. Para definir o polimorfismo genético da população foram genotipados em eletroforese capilar 15 *loci* microssatélites, usando uma plataforma de sequenciador automatizado em sistema multiplex. O primeiro capítulo do estudo apresenta e discute os métodos usados para caracterizar, detectar e validar o polimorfismo microssatélite da população. Os resultados detectaram erros de tipagem alélica e permitiram concluir que a melhor forma de detectá-los foi obtida com o estudo da segregação alélica de cada *locus* dentro das progênies. Esse estudo também permitiu a inferência dos genótipos de 120 ancestrais maternos. O capítulo seguinte apresenta o resultado

das análises de agrupamento *multilocus* usando algoritmo bayesiano para definir o número mais provável de grupos genéticos existentes na população selecionada e os parâmetros genéticos que melhor definem a estrutura genética. Essas análises foram efetuadas sob dois critérios: admitindo-se equilíbrio de Hardy-Weinberg e admitindo-se endogamia. Também foram utilizadas estatísticas *ad hoc* para determinar o K mais provável. Concluiu-se ao final do estudo que existam cinco grupos genéticos prováveis, os quais não obtiveram a coincidência esperada com às chamadas Florestas Plantadas (FPs) – plantações onde foram selecionados os indivíduos que geraram a população selecionada. O quê os resultados sugerem é que esses cinco grupos foram gerados pela pressão de seleção dentro de progênies para características de volume de madeira produzida, e ainda pela ocorrência de acasalamentos preferenciais entre três origens genéticas distintas. Os valores de endogamia F_{ST} (coeficiente de endogamia de Wright) detectados foram de 0,194 a 0,384, e são relativamente superiores ao que seria esperado para uma população inicial de seleção de uma espécie de polinização aberta. Esses resultados, apesar de terem como alvo o preparo dessa população para um estudo de genética de associação, podem ser de grande valor prático para os próprios programas de melhoramento, uma vez que o conhecimento da dinâmica da estrutura genética permite monitorar o nível de endogamia a cada geração de seleção e assim, prolongar a vida útil do programa de melhoramento.

Palavras-chave: Genética de Associação, Estrutura Genética, Microssatélites, Parâmetros Genéticos, Inferência Bayesiana.

ABSTRACT

Fast-growing wood species supply in a sustainable way the demand for woody biomass, reducing the pressure on natural forests. However, reforestations with pine have been producing young wood with low quality. Thus, the genetical improvement for traits related to wood quality or others complex and of difficult measurement, becomes a priority to almost all forest improvement programs. Accordingly, high expectations have been directed to studies of genetic association using candidate genes, since they provide the possibility to estimate genetic and phenotypic correlation between polymorphic alleles and traits of interest to be improved. Such a strategy would be a way of predicting genetic gains from the implementation of a program based on marker-assisted selection. However, the existence of spurious genetic structure is reported in the literature as a major source of statistical bias in studies of genetic association. It is known that a major challenge, even with the advancement of all genomic and analytic technologies, is to define what are the real genetic groups existing within a given population. To define the genetic groups in an artificial population becomes even more challenging, as it is not known for sure the historical origin of the individuals that were used to generate these populations. The present research assumes that setting the most probable number of K (or distinct genetic groups) and correctly assigning the individuals to their clusters is the most efficient strategy to avoid the statistical bias in the genetic association mapping. Setting out from this hypothesis, the objective was to characterize the genetic structure of a breeding population of *Pinus taeda*, searching for the true K or the most likely, aiming to prepare the population for a future association genetic study. The population consists of 1,130 individuals at 12 years of age, representing 120 open-pollinated progenies, which were previously selected within progenies for wood volume production. In order to define the genetic polymorphism of the population, 15 microsatellite *locus* were genotyped by a multiplex system using capillary electrophorese in an automated sequencer platform. The first chapter of the study presents and discusses the strategies implemented to characterize, detect and validate the microsatellite polymorphism of the target population. Genotyping errors were detected in the results. It was concluded that the most efficient method to detect these errors is the study of the *locus* segregation within the family. This study also provided the inference of genotypes of 120 maternal ancestral. The following chapter presents the result of the *multilocus* bayesian clustering analysis to define the most probable K of the breeding population, and the genetic parameters that best describe the genetic structure. These

analyses were implemented under two criteria: assuming Hardy-Weinberg equilibrium and assuming inbreeding. Also, it was used two *ad hoc* statistics to define the most probable K. At the end of the study it was concluded that there are five probable genetic groups, which unexpectedly do not match to the seed production places – called Forestry Plantations (FPs) – where the individuals were selected in order to form the breeding population. Apparently, these five groups were generated by the selection pressure for wood volume improvement that occurred within progenies; also by the occurrence of assortative matings between three different genetic backgrounds. The inbreeding F_{ST} values (Wright's inbreeding coefficient) detected were 0,194 to 0,384, which are relatively higher than it would be expected for a population of initial selection from an open-pollinated species. These results, though aiming to prepare this population for a genetic association study, may be of great practical value to their own breeding programs, since the knowledge of the dynamics of genetic structure allows the monitoring of the level of inbreeding in each generation of selection, and thus to prolong the existence of the breeding program.

Key Words: Genetic Association, Genetic Structure, Microsatellite Markers, Genetic Parameters, Bayesian Inference.

1. INTRODUÇÃO

Uma das maiores metas dos programas de melhoramento florestal atualmente, em função do longo intervalo entre gerações de seleção, é a definição de estratégias que tornem a Seleção Assistida por Marcadores (SAM) uma prática tangível para características quantitativas e complexas. São essas, pois, que representam a maioria das características de interesse econômico do setor florestal.

Para que a SAM se torne uma realidade, é necessário que uma proporção do controle genético da característica quantitativa possa ser dissecada. Dentre as estratégias que vêm demonstrando esse potencial, maior enfoque é dado aos estudos de mapeamento de associação para determinar a correlação existente entre polimorfismos e fenótipos de interesse. De modo simplista, tais resultados permitiriam predizer quais seriam os ganhos obtidos com a seleção das características, ainda em estágios iniciais das populações de melhoramento.

Contudo, apesar de promissores, um dos maiores entraves para esses estudos reside na existência de estruturação (ou fragmentação) genética espúria e/ou não detectada na população de mapeamento. Tal fenômeno é apontado na literatura como uma das maiores fontes de viés estatístico das análises de associação.

Estudos de Genética de Populações se concentram na análise da variação da freqüência alélica entre e dentro de populações. Dos índices genéticos mais utilizados para definir a estrutura populacional, destaque é dado ao coeficiente de endogamia de Wright (F_{ST}). No entanto, para que não haja distorção na estimação desses índices genéticos é primordial que seja definido quais indivíduos representam grupos genéticos distintos dentro de uma população, pois a variância da freqüência alélica é estimada usando os genótipos dos indivíduos que compõem cada grupo genético.

Mesmo com os maiores avanços nos algoritmos analíticos e na tecnologia genômica, ainda é desafiador definir e apontar qual ou quais são os grupos genéticos distintos existentes que estejam causando sinais de fragmentação em uma população. Ainda, há a dificuldade de definir se esses sinais são reais, pois podem ser gerados de outros fatores e não da existência de fragmentação em si. O problema torna-se ainda mais desafiador quando a população em questão é formada por um conjunto de indivíduos para os quais poucas informações a respeito

das origens são disponíveis, e que na verdade fazem parte da população por terem sido forçosamente selecionados para características de interesse e não por forças evolutivas naturais.

Surge então o seguinte questionamento: “Como detectar a existência de fragmentação genética em população artificial de melhoramento para obter êxito em estudos de mapeamento de associação?” As respostas para essa pergunta são escassas e, quando existentes, são bastante fragmentadas na literatura.

Entre os métodos de agrupamento *multilocus* existentes que são usados para determinar e descriminar os grupos genéticos, o método de agrupamento bayesiano disponibilizado pelo *software* STRUCTURE (PRITCHARD *et al.*, 2000) tem sido o mais utilizado nas publicações recentes.

Nesse método, o algoritmo bayesiano do modelo assume a existência de K populações ou grupos genéticos e cada um desses, possui uma freqüência alélica característica para cada *locus*. Como o número de K é desconhecido, o modelo utiliza uma distribuição de probabilidade *a priori* [Pr(K | X)] e infere K, em função da distribuição *a posteriori* [Pr(X | K)]. Os indivíduos da amostra analisada são, então, distribuídos probabilisticamente para cada K, ou são atribuídos para mais de um K, caso os genótipos indiquem que sejam híbridos entre os grupos.

GAO *et al.* (2007) chamam atenção para o fato de que a endogamia pode induzir a sinais espúrios de fragmentação quando a população é analisada segundo o critério de equilíbrio de Hardy-Weinberg (HW), conforme é o modelo assumido nas análises do STRUCTURE. Os autores disponibilizam o software INSTRUCT baseado em um modelo que admite endogamia em análises de agrupamento *multilocus*, que permite estimar as freqüências genotípicas com base nas taxas de endogamia e de endocruzamento. Os mesmos autores confirmam, com simulações coalescentes, que o método é eficiente em corrigir o viés de sinais espúrios de fragmentação.

Como populações de plantas apresentam estruturas genéticas geralmente tidas como de difícil resolução, principalmente por ser comum a ocorrência de acasalamento entre parentes e a hibridização de diferentes origens (CAMUS-KULANDAIVELU *et al.*, 2007), o método baseado em endogamia torna-se relevante.

Um marcador amplamente empregado nessas análises de agrupamento é o marcador microssatélite, ou SSR (*Simple Sequence Repeats*), por ser considerado seletivamente neutro, codominante e pelo grau de polimorfismo que possui em potencial. Apesar dos marcadores microssatélites serem promissores em estudos genéticos, erros de genotipagem reduzem drasticamente a confiabilidade de estudos genéticos.

O processo de genotipagem como um todo, envolve uma série de atividades. Essas se iniciam com o isolamento do DNA genômico das amostras, a amplificação dos segmentos microssatélites em PCR (*Polymerase Chain Reaction*), a diluição dos produtos de PCR e o preparo da placa para entrada na plataforma de genotipagem; que por sua vez, envolve a diluição e homogenização do marcador de banda padrão com os fragmentos amplificados de DNA e a denaturação do DNA amplificado. Assim, vários momentos em que as manipulações humanas ocorrem, fazem parte dessas atividades, o que torna o processo sujeito a erros de inúmeras naturezas. Erros de genotipagem podem ocorrer por várias causas, mas a sua existência e efeito podem ser limitados se as causas na produção e análise dos dados forem levadas em consideração.

O presente estudo parte da hipótese de que, validando o polimorfismo microssatélite, definindo o número de grupos genéticos mais prováveis (K: grupos genéticos distintos) e então, apontando como os indivíduos da população se distribuem dentro desses Ks, seja a alternativa mais eficiente para conter o viés estatístico em estudo de mapeamento de associação.

A população-alvo da pesquisa é constituída por 1130 indivíduos com 12 anos de idade, os quais representam 120 famílias de polinização aberta, que faz parte de um amplo programa de melhoramento da espécie do setor privado no estado de Santa Catarina, Brasil.

O primeiro capítulo desta pesquisa apresenta e discute os métodos utilizados e os resultados encontrados para caracterizar, detectar e validar o polimorfismo microssatélite gerado. Tais resultados são relevantes, uma vez que mostram que uma cautelosa tipagem alélica é necessária, em função de que a genotipagem em larga escala e em sistema multiplex estejam mais propensos à ocorrência de erros de naturezas diversas. Aproveitar a estrutura familiar da população também permite analisar a segregação de *locus* dentro de famílias e inferir os genótipos maternos. Esse procedimento possibilita detectar erros de genotipagem e presença de alelos nulos, que possam passar despercebidos e inviabilizar estudos genéticos.

O capítulo seguinte apresenta o resultado de análises de agrupamento *multilocus* usando algoritmo bayesiano sob dois critérios: admitindo equilíbrio de Hardy-Weinberg (programa estatístico STRUCTURE – PRITCHARD *et al.,* 2000) e admitindo endogamia (programa estatístico INSTRUCT – GAO *et al.*, 2007), para definir qual o número mais provável de grupos genéticos existentes e quais os parâmetros genéticos que melhor definem a estrutura genética. Os resultados dessas análises demonstram o contraste entre essas duas abordagens, mostrando o caminho para tentar resolver estruturas genéticas complexas pouco referenciadas na literatura.

A contribuição teórica do conhecimento metodológico aplicado torna promissor o presente estudo, uma vez que propõe uma estratégia que possibilita validar uma resolução que melhor define a estrutura genética de populações complexas. A relevância dos resultados obtidos e a forma com que foram explorados, são de interesse tanto para estudos voltados à Genética de Populações, quanto para os programas de melhoramento, por auxiliarem no monitoramento da variabilidade genética da população selecionada.

2. OBJETIVO GERAL

Caracterizar a estrutura genética de uma população selecionada de 120 progênies de polinização aberta da espécie *Pinus taeda* L., constituída por 1130 indivíduos utilizando 15 marcadores microssatélites, com o propósito de prepará-la para um estudo de mapeamento de associação.

2.1 Objetivos Específicos

- Validar a fidedignidade dos alelos detectados no procedimento de genotipagem, assumidos como indicadores do polimorfismo dos 15 *loci* estudados para a população.

- Comparar dois modelos de análises bayesianas de detecção do grau de fragmentação da população: assumindo equilíbrio de Hardy-Weinberg e assumindo endogamia.

- Definir o número mais provável de grupos genéticos existentes, pelo apontamento da distribuição dos indivíduos dentro e entre os grupos genéticos e pela inferência dos parâmetros genéticos coeficiente de endogamia de Wright (F_{ST}) e heterozigosidade esperada (He) de cada grupo genético.

3. REVISÃO BIBLIOGRÁFICA

3.1 Melhoramento Genético do *Pinus taeda* L.

A espécie *Pinus taeda* L. pertence ao reino *Plantae*, ao grupo das *Gimnospermae*, à divisão *Pinophyta*, à classe *Pinopsida*, à ordem *Pinales* e à família *Pinaceae*. Vulgarmente conhecida como *pinus* ou pinheiro, é originária do Novo Mundo e faz parte do grupo de espécies distribuídas no Canadá e Estados Unidos da América (GOLFARI, 1971).

É uma espécie alógama com elevada taxa de fertilização cruzada, organismo diplóide, sendo 12 o número básico de cromossomos. Estima-se que o conteúdo haplóide do genoma seja em torno de 20,000 Mbp (WAKAMIYA *et al.*, 1993).

As gimnospermas possuem um dos maiores e mais complexos genomas entre todos os organismos vivos. O estudo desses genomas representa uma fonte inesgotável de informação para estudos evolutivos, uma vez que forças evolutivas modelaram o genoma desse gênero há pelo menos 87–193 milhões de anos, período em que o gênero divergiu do *Picea* (GROTKOPP *et al.*, 2004).

A produção florestal sustentável no Brasil, referente ao plantio de pinus e eucalipto, atingiu cerca de 191,4 milhões de m³/ano. A produção de madeira em toras de pinus concentra-se nas regiões Sudeste e Sul, com 46,5 milhões de m³/ano da produção sustentável nacional. Esta concentração resulta do desenvolvimento da indústria madeireira, especialmente na produção de madeira serrada, compensados e painéis reconstituídos na região Sul do país. Em termos econômicos, a contribuição do setor florestal na formação do produto interno bruto nacional é de 3,4% (SBS, 2008).

Em 2007, a área total de florestas plantadas apresentava uma extensão de 5,98 milhões de hectares (ha) no país, e a área com *Pinus taeda* nos estados do sul representavam 1,8 milhões de ha (SBS, 2008).

O período em que ocorreram as primeiras introduções da espécie no Brasil, foi em 1948 (SHIMIZU, 2011). Até há pouco tempo, a maioria dos programas de melhoramento genético se concentravam especificamente no aumento da produtividade de madeira por área. Existem informações informais, porém recorrentes entre os melhoristas florestais, de que tais incrementos se encontram hoje, em torno de 30 a 40% para o volume de madeira por área, em relação aos obtidos com as primeiras florestas plantadas no País.

Durante encontros com melhoristas florestais, foi observado que ocorrem inúmeras diferenças e singularidades nos métodos de melhoramento que são adotados. No entanto, esses possuem em comum alguns procedimentos básicos para obter ganhos genéticos a cada geração. Tais procedimentos se resumem em constituir uma população base, selecionar e recombinar os indivíduos selecionados, testar o desempenho dos recombinantes em delineamentos com repetições para, finalmente, selecionar os melhores indivíduos que constituirão uma nova população de seleção. Para realizar seleção genética e aumentar a eficiência de ganhos a cada ciclo seletivo, os indivíduos da população são selecionados a partir do desempenho em teste de progênies. Tais progênies devem ser instaladas em delineamentos experimentais com repetições, e podem ser tanto de polinização controlada (*full-sib*), quanto de polinização aberta (*open-sib*) – em que somente o lado materno é controlado.

A formação da população base, segundo SHELBOURNE (1973), deve ser efetuada a partir da seleção individual e recomenda uma área mínima de 120 hectares (ha) para que essa seleção seja respeitada, e/ou que nessa área deva existir cerca de 100 mil indivíduos. Quanto ao tamanho da população para reprodução, BURDON e SHELBOURNE (1973) citam que é muito arbitrário fixar um número preciso de árvores a serem selecionadas para compor essa população. Segundo os autores, deve-se pensar em uma quantidade próxima a 200 ou mais árvores para formar a população base de alguns programas.

ZOBEL *et al.* (1973) mencionam que os pomares de gerações avançadas devem ser constituídos por árvores com genótipo conhecido e, para tanto, há a necessidade de sintetizar populações de base genética ampla, formadas a partir da produção inicial do pomar. Tal procedimento poderia garantir um número de clones não aparentados suficientemente grande, o que ajudaria a minimizar os problemas de endogamia. Porém, ressaltam que uma população tão grande quanto 300 clones ainda não preveniriam a ocorrência de acasalamentos entre parentes em gerações de melhoramento mais avançadas.

3.1.1 Seleção Assistida por Marcadores

Quando a característica-alvo do programa de melhoramento é de difícil mensuração (caracteres complexos), ou de longo período para se expressar, o melhorista depara-se então com um grande obstáculo: o prolongado tempo necessário para atingir ganhos com a seleção. Isso explica o forte interesse e os incentivos governamentais que são destinados às pesquisas

que focam na Seleção Assistida por Marcadores (SAM), uma vez que representariam uma possibilidade de selecionar ao nível genômico, ainda em fases iniciais de desenvolvimento da população (FAO, 2007).

O maior desafio do melhoramento genético de planta, atualmente, consiste na identificação de marcadores genéticos que controlam as características de interesse, de modo que possam efetivamente ser utilizados em programas de seleção assistida por marcadores. Tais programas, apesar de terem sido vislumbrados há mais de duas décadas, tiveram poucas aplicações práticas, uma vez que a maioria das características de interesse possui uma natureza quantitativa.

Análises de Ligação e o Mapeamento de Associação são as ferramentas mais usadas para dissecar a variação fenotípica de caracteres quantitativos e complexos, os quais representam a maioria de importância econômica para inúmeras espécies (ZHU *et al.*, 2008).

Contudo, dissecar uma característica quantitativa por completo ainda está muito distante de acontecer, conforme discussão em ZHU *et al.* (2009). Os autores ressaltam os pontos que estão travando o avanço desse conhecimento e argumentam que é crucial a existência de uma maior interligação entre as diversas disciplinas da biologia, da genética e da biometria. Propõem uma alternativa para estudar as características quantitativas na presente era da Biologia dos Sistemas (*Systems Biology Era*), que passaria a uma nova era, a qual chamam de Genética Quantitativa de Sistemas (*Systems Quantitative Genetics*).

3.1.2 Perspectivas do Mapeamento de Associação

Existe um forte interesse na utilização de Mapeamento de Associação para a identificação de genes responsáveis por variações quantitativas de caracteres complexos, relevantes tanto para produção quanto para aspectos evolutivos. Avanços recentes na tecnologia genômica e no desenvolvimento de métodos de análises estatísticas robustas, tornam o mapeamento de associação palpável e passível de ser aplicado em programas de pesquisas com plantas (ZHU *et al.*, 2008).

O Mapeamento de Associação também é referenciado como Mapeamento de Desequilíbrio de Ligação. Este, por sua vez, é dividido em duas grandes categorias: o mapeamento que usa a estratégia de aproximação via genes candidatos e o de ampla cobertura genômica (ZHU *et al.*, 2008).

As pesquisas nessa categoria de estudo estão demonstrando inúmeras vantagens em relação às análises de ligação: maior poder de recombinação (várias gerações de cruzamento);

maior variabilidade genética (maior número de alelos analisados), maior precisão nas estimativas do desequilíbrio de ligação (DL), entre outras. A genética de associação aliada à estratégia da abordagem via genes candidatos diretamente envolvidos na expressão de características complexas, representa uma metodologia recente e promissora, gerando um mapeamento refinado para identificar polimorfismos em regiões gênicas, que têm uma elevada probabilidade de estarem envolvidos com a variabilidade na característica de interesse (YU *et al.*, 2002, 2006, 2009).

Nesse contexto, pesquisas internacionais (GONZÁLEZ-MARTÍNEZ *et al.*, 2006a e 2006b; POT *et al.*, 2006), vêm demonstrando a possibilidade de se utilizar marcadores gênicos para a seleção das plantas, por meio da detecção de associação estatística entre alelos de genes candidatos com a variabilidade do fenótipo de interesse de características complexas. Assim, o objetivo da genética de associação, em poucas palavras, busca identificar polimorfismos ao nível gênico e analisar quais desses representam marcadores relevantes para uma possível SAM.

BROWN *et al.* (2003) realizaram experimentos para mapear QTLs (*Quantitative Trait Loci*) envolvidos com propriedades da madeira em *P. taeda*, em múltiplas origens genéticas, ambientes e estágios de desenvolvimento, para gerar estimativas mais precisas da magnitude dos efeitos do QTL e predizer a expressão do QTL em um dado estágio de desenvolvimento ou em um dado ambiente particular.

POT *et al.* (2006), visando identificar regiões genômicas envolvidas na variabilidade de propriedades físicas e químicas de *Pinus pinaster*, utilizaram uma população de terceira geração de cruzamento controlado composta por 186 indivíduos. Detectaram um total de 54 QTLs, com uma média de 2,4 QTLs para cada caráter considerado separadamente. Encontraram *clusters* para esses QTLs em muitos pontos do genoma, sugerindo a existência de efeitos pleiotrópicos de um número limitado de genes.

GONZÁLEZ-MARTÍNEZ *et al.* (2006a) estudaram os padrões de polimorfismo de 18 genes candidatos relacionados ao estresse hídrico em *P. taeda*. Os dados analisados, baseados em um conjunto de 21 marcadores de microssatélites (SSR) neutros, não demonstraram nenhum desvio da neutralidade para estrutura genética da população amostrada.

Em um segundo experimento GONZÁLEZ-MARTÍNEZ *et al.* (2006b) realizaram um estudo de genética de associação para o caráter qualidade de madeira em *P. taeda* e ressaltaram que esse é um método poderoso para dissecar caracteres adaptativos complexos, pois permitiu o mapeamento em alta resolução de ampla cobertura para a variação fenotípica

e genotípica detectadas dentro de um único experimento. Nesse trabalho, utilizaram 58 marcadores de polimorfismo único (SNP – *Single Nucleotide Polymorphism*) de 20 genes candidatos, alguns desses foram efetivamente estudados em seu primeiro experimento e os outros faziam parte de estudos prévios de BROWN *et al.* (2003).

GONZÁLEZ-MARTÍNEZ *et al.* (2006b) detectaram que ocorria um rápido decréscimo de DL nas regiões intragênicas de coníferas e, em razão disso, concluíram que os SNPs que apresentaram algum nível de associação genética provavelmente estão localizados de modo muito próximos aos polimorfismos causais. Esse estudo de genética de associação multigênico em espécies florestais, demonstrou a possibilidade plausível de se utilizar a estratégia de genes candidatos para dissecar caracteres adaptativos complexos. Os pesquisadores também demonstraram que se trata de uma metodologia particularmente útil para espécies, tais como coníferas, em que as estratégias que levam em conta toda a região genômica são limitadas pelo imenso genoma dessas espécies.

O nível de DL detectado para as regiões intragênicas no experimento de GONZÁLEZ-MARTÍNEZ *et al.* (2006a) para populações de diferentes regiões de *Pinus taeda* foi r^2 (*average pairwise*) de 0,30, decrescendo significativamente (0,50 para 0,20) em segmentos maiores do que 800 bp (pares de bases). Da mesma forma não detectaram nenhum DL aparente entre regiões intergênicas, sugerindo hipótese de que tenha ocorrido um processo de expansão populacional. Embora ainda não existam informações conclusivas a respeito das extensões existentes de DL para as espécies florestais, sabe-se que as variações que podem ocorrer em populações naturais de espécies florestais são modeladas em conseqüência da interação de vários efeitos genéticos, como flutuação gênica, migração, deriva genética, estrutura populacional e seleção natural.

Não obstante a essas vantagens, cabe ressaltar que um pré-requisito para iniciar um estudo de associação baseado em desequilíbrio de ligação, é a realização de prévio estudo para caracterizar a estrutura genética da população baseado em marcadores neutros e não ligados. O efeito da estratificação de populações é a causa mais comum de ocorrência de viés estatístico em estudos de associação (BUCKLER e THORNSBERRY, 2002; HIRSCHHORN e DALY, 2005).

O desenvolvimento de teorias, simulações de computador e as evidências empíricas dos estudos de população, continuam a indicar que a estratificação da população, devido à mistura de diferentes origens genéticas, assim como desvios de panmixia, podem distorcer os estudos de associação genética e produzir resultados falso-positivos (HALDER e SHRIVER, 2003).

3.2 Estrutura Genética de Populações

A ciência Genética de Populações define uma população pela frequência de alelos que a caracterizam, e analisa o comportamento da distribuição destas frequências sob o efeito de processos evolutivos. Havendo assim a necessidade de existir mais de uma geração de indivíduos, no estudo de uma dada população de referência. Caso as frequências alélicas não sofram alterações com o decorrer de gerações, considera-se que a população esteja em equilíbrio de Hardy-Weinberg. Entretanto, para que este equilíbrio possa ocorrer, certas premissas devem existir, como: a ausência de seleção natural, mutação, e de migração, e tamanhos populacionais infinitamente grandes (CROW, 1999).

Caso seja detectada uma sistemática diferenciação na distribuição das frequências alélicas entre os indivíduos que formam a população estudada, resultando em agrupamentos não aleatórios, diz-se que a população está estratificada ou estruturada. Notou-se todavia, que a terminologia encontrada na literatura é ambígua. Assim, este estudo considera que estrutura genética de uma população seja a frequência de alelos que a definem. Já, considerar a população estruturada é o mesmo que fragmentada ou estratificada, ou fora do equilíbrio de Hardy-Weinberg.

Após extensa revisão de literatura, WAPLES e GAGGIOTTI (2006) concluíram que as definições para conceituar uma população tornam um ponto evidente: "não existe uma única resposta correta para a questão: O que é uma população?”. A natureza flexível desses conceitos, implica que diferentes terminologias possam ser aplicadas a uma ampla gama de situações com que se deparam ecologistas e biólogos evolucionistas, e promove ambigüidade e confusão entre a comunidade científica.

Os autores partem de uma situação hipotética utilizando diferentes modelos de metapopulação, e levantam questões fundamentais no que diz respeito à identificação de uma população. Utilizando duas situações extremas, na primeira existe um isolamento completo entre fragmentos ou subpopulações, e na segunda ocorre acasalamento aleatório entre todos os indivíduos de todas as subpopulações. Concluem que: no primeiro cenário os fragmentos deveriam ser considerados populações individuais; já no segundo cenário as subpopulações passam a ser arbitrárias. Contudo, em situações intermediárias a esses dois extremos, que é o que normalmente ocorre em populações naturais, os autores questionam qual é o ponto dentro de um gradiente de diferenciação genética, em que as populações se tornam tão diferenciadas que deveriam ser consideradas como entidades distintas.

3.2.1 Populações Fragmentadas

Os impactos da fragmentação das populações sobre a diversidade genética, diferenciação e endogamia, dependem do nível de fluxo gênico entre os fragmentos populacionais. Este, por sua vez, depende do número, do tamanho, da distância e da distribuição geográfica dos fragmentos populacionais e também da habilidade de dispersão gamética das espécies. A endogamia resultante da fragmentação populacional pode ser usada para medir o grau de diferenciação que ocorreu entre os fragmentos. A diferenciação entre os fragmentos ou subpopulações está diretamente relacionada com os coeficientes de endogamia dentro e entre populações (FRANKHAM *et al.*, 2008).

WRIGHT (1951) e MALÉCOT (1948) apresentaram a Estatística F como uma ferramenta para descrever a partição da diversidade genética dentro e entre populações. Em seu trabalho, WRIGHT (1931) demonstra que a quantidade de diferenciação genética entre as populações tem uma relação previsível às taxas de importantes processos evolutivos (migração, mutação e deriva). Grandes populações, entre as quais ocorre muita migração, por exemplo, tendem a ser pouco diferenciadas. Por outro lado, populações pequenas, entre as quais há pouca migração, tendem a ser bastante diferenciadas.

A endogamia na população total (F_{IT}) pode ser dividida dentro dela devido à endogamia dos indivíduos em relação às suas subpopulações ou fragmentos (F_{IS}), e a endogamia devido à diferenciação entre subpopulações, em relação à população total (F_{ST}) (FRANKHAM *et al.*, 2008).

Mas F_{ST} é mais do que uma estatística descritiva e uma medida de diferenciação genética. A F_{ST} está diretamente relacionada à variação na freqüência dos alelos entre populações e inversamente ao grau de semelhança entre os indivíduos dentro das populações. Se F_{ST} for pequeno, isso significa que as freqüências alélicas em cada população são muito semelhantes; se for grande, que as freqüências alélicas são muito diferentes. Se, a seleção natural favorece um alelo em detrimento de outros, em um *locus* específico em algumas populações, a F_{ST} para aquele *locus* será maior do que para *loci* em que as diferenças entre as populações são puramente resultado de deriva genética. Assim, as varreduras genômicas que comparam as estimativas de F_{ST} de um único *locus* com as de ampla cobertura genômica, podem identificar regiões do genoma que foram submetidos à seleção diversificadora (HOLSINGER e WEIR, 2009).

WRIGHT (1951) introduziu F_{ST} como um dos três parâmetros inter-relacionados, utilizados para descrever a estrutura genética de populações diplóides: F_{IT}, mede a correlação

entre os gametas de um indivíduo em relação a toda a população; F_{IS}, a correlação entre gametas de um indivíduo em relação à subpopulação em que ocorrem; e, F_{ST}, a correlação entre gametas escolhidos aleatoriamente dentro de uma mesma subpopulação em relação a toda a população.

O coeficiente de endogamia dentro da população (*f)* é um termo que pode ser confuso (HOLSINGER e WEIR, 2009). Na prática, *f* é uma medida da freqüência de heterozigotos, comparada ao que é esperado quando os genótipos estão nas proporções de equilíbrio de Hardy-Weinberg. A endogamia leva a uma deficiência de heterozigotos em relação às expectativas de Hardy-Weinberg e nesse caso *f* será positivo. Mas, se os indivíduos evitam endogamia ou se há vantagem para o heterozigoto, os heterozigotos serão mais freqüentes do que o esperado no equilíbrio de Hardy-Weinberg e os valores de *f* serão negativos. Em suma, os valores de *f* são uma medida de quão diferentes são as proporções de genótipos dentro das populações em relação às expectativas de Hardy-Weinberg, com valores positivos, indicando uma deficiência de heterozigotos e valores negativos, indicando um excesso.

Assim, a Estatística F de Wright e especialmente o F_{ST}, fornecem importante conhecimento sobre os processos evolutivos que influenciam a estrutura da variação genética dentro e entre populações, e esses parâmetros estão entre os mais utilizados para descrever populações (HOLSINGER e WEIR, 2009).

3.2.2 Polimorfismos Genéticos para Estudos Genéticos de Populações

Os marcadores microssatélites (SSR: *Simple Sequence Repeats*) têm sido os marcadores de DNA mais amplamente utilizados para caracterizar coleções de germoplasmas, devido à sua facilidade de manejo, custo relativamente baixo e alto grau de polimorfismo proveniente de um grande número de alelos por *locus* (VIGNAL *et al.*, 2002; GONZÁLEZ-MARTÍNEZ *et al.*, 2006b e HOLSINGER e WEIR, 2009).

Recentemente os marcadores SNPs (*Single Nucleotide Polymorphism*) receberam grande atenção, porque ocorrem com freqüência muito maior no genoma do que os SSRs e a sua genotipagem pode ser facilmente automatizada (VAN INGHELANDT *et al.*, 2010). Contudo em função do potencial polimórfico, esses autores sugerem que sejam necessários entre 7 a 11 vezes mais SNPs do que SSRs, quando o objetivo reside em analisar estrutura genética e a diversidade genética da população.

Embora os microssatélites apresentem uma série de vantagens, a literatura aponta que erros de genotipagem não são incomuns, tanto para esses marcadores, como para outros marcadores. Erros de genotipagem ocorrem quando um genótipo determinado para o indivíduo não é verdadeiro. Tais erros, se não levados em consideração, podem afetar a confiabilidade das inferências biológicas concluídas nos estudos genéticos (THOMPSON, 1976; POMPANON et al., 2005; HOFFMAN e AMOS, 2005; BONIN et al., 2004; GAGNEUX et al., 1997a; GAGNEUX et al., 1997b).

Os marcadores moleculares permitem estimar o grau de parentesco entre os membros de grupo de germoplasma, mesmo quando seu histórico de seleção já não está disponível ou quando se tenha tornado demasiado complexo para uma análise clássica de *pedigree*. O campo da genética de populações tem vários procedimentos para a estimação de parâmetros ao seu dispor, mas quando os indivíduos genotipados possuem forte endogamia, a aplicação destes não cumpre uma série de pressupostos teóricos, para o qual os estimadores foram construídos (MAENHOUT *et al.*, 2009).

3.2.3 Estimativas de Parâmetros em Populações Estruturadas

O interesse na estruturação genética de populações naturais surgiu no início dos anos 30, após a publicação da síntese de genética de populações por FISHER (1930) e WRIGHT (1931), quando foi revelada a divergência de opiniões a respeito do efeito da estrutura da população na dinâmica da evolução por seleção natural.

Informações sobre a diversidade genética e sobre a estrutura populacional em materiais elite de melhoramento são de fundamental importância para o incremento das culturas (HALLAUER e MIRANDA, 1988). Vários caminhos têm sido sugeridos na literatura para atingir esse objetivo. No entanto, essas estimativas muitas vezes podem ser inviáveis ou duvidosas, principalmente para espécies alógamas, em função da dificuldade de controlar o *pedigree* (LÜBBERSTEDT *et al.*, 2000).

Os estatísticos têm desenvolvido várias abordagens diferentes para estimar os parâmetros de dados. Três abordagens amplamente utilizadas são: o método dos momentos, o método de máxima verossimilhança e o método de inferência bayesiana (HOLSINGER e WEIR, 2009).

Para a Estatística F, as estimativas realizadas pelo método dos momentos (WEIR e COCKERHAM, 1984) são baseadas em análises de variâncias de freqüências alélicas, e a Estatística F é definida em termos da decomposição dos componentes de variância. De modo

simplificado, se a variância entre as populações é a mesma do que a variância dentro, então a população não está estruturada.

Estimativas bayesianas apresentam muitas das vantagens associadas às estimativas de máxima verossimilhança, uma vez que estas empregam a mesma verossimilhança para relacionar os dados com parâmetros desconhecidos. Porém, diferem das estimativas de máxima verossimilhança porque a probabilidade é modificada pela introdução de distribuições *a priori*, sobre os parâmetros desconhecidos e as estimativas são baseadas na distribuição *a posteriori*, o que é proporcional ao produto da probabilidade das distribuições *a priori*.

Ambas, a máxima verossimilhança e a inferência bayesiana, sofrem a desvantagem de que simples expressões algébricas para as estimativas, raramente podem ser efetuadas. Em vez disso, as estimativas são obtidas por métodos computacionais extremamente exigentes.

Como os métodos MCMC (*Markov Chain Monte Carlo*) de amostragem, utilizados para análise bayesiana de modelos, não exigem que um único ponto de máxima verossimilhança seja identificado, as estimativas bayesianas podem ser inferidas até mesmo para modelos complexos, com milhares ou dezenas de milhares de parâmetros, em que a maximização numérica da probabilidade seria difícil ou impossível (GELFAND e SMITH, 1990).

Os métodos bayesianos permitem que inferências sejam realizadas para as Estatísticas F e extensões permitem relacionar a Estatística F com variáveis demográficas, com covariâncias ambientais, explorado dentro de um contexto de modelo único (SAMANTA *et al.*, 2009), embora tais implementações sejam computacionalmente exigentes (GAGGIOTTI *et al.*, 2002; HUBISZ *et al*., 2009).

CAPÍTULO I

DETECÇÃO E VALIDAÇÃO DO POLIMORFISMO MICROSSATÉLITE PARA A CARACTERIZAÇÃO GENÉTICA DE UMA POPULAÇÃO SELECIONADA DE *Pinus taeda* L.*

RESUMO

No presente trabalho, marcadores microssatélites foram genotipados em uma população de 120 progênies de polinização aberta, que pertencem a um programa de melhoramento do setor privado, instalado no sul do Brasil. Quinze *loci* microssatélites marcados com fluorescência foram caracterizados em eletroforese capilar automatizada (MegaBace 1000 Fragment Analyzer GE). Apesar da plataforma de genotipagem ter demonstrado um potencial muito promissor, percebeu-se que muitos erros de genotipagem podem comprometer os dados e invalidar estudos para essa população. Erros de genotipagem afetam a maioria dos dados biológicos e podem influenciar sensivelmente as conclusões de estudos genéticos. Mesmo assim, eles são muitas vezes negligenciados. Erros podem ocorrer por várias causas, mas a sua existência e efeito podem ser limitados se as causas na produção e análise dos dados forem levadas em consideração. Assim, as tipagens alélicas sofreram cautelosa validação. Alelos nulos foram detectados em 10 *loci*, representando 32,7% do total de alelos. Dez *locos* não segregaram em proporções mendelianas. Embora alguns procedimentos tenham sido desenvolvidos para tentar conter erros de genotipagem, foi verificado que para o sistema de genotipagem multiplex e para o volume de amostras envolvidas, nenhum deles foi eficaz. Aqui é relatado o protocolo utilizado na tentativa de conter as principais causas de erros, de modo a torná-los independentes. Concluiu-se que a plataforma automatizada de eletroforese capilar baseada em fluorescência é um método rápido e eficaz para genotipagem de microssatélites em larga escala, contanto que cautelosa análise da tipagem alélica seja efetuada para conter os freqüentes erros detectados em sistema multiplex.

Palavras-chave: *Pinus taeda*; microssatélite; alelos nulos; erros de genotipagem; genética de populações

* A configuração do capítulo segue as exigências de formatação da revista "Plant Methods", (http://www.plantmethods.com/).

DETECTION AND VALIDATION OF MICROSATELLITES FOR THE CHARACTERIZATION OF A *Pinus taeda* L. BREEDING POPULATION

ABSTRACT

Background

Genotyping errors affect most data and can markedly influence the biological conclusions of a study; even so, they are oftentimes neglected. Errors may occur due to several causes, but their occurrence and effect can be limited if these causes in the production and the analysis of data were considered.

Results

Microsatellites were screened in a 120 open-sib breeding population of *Pinus taeda* from a private program of southern Brazil. Fifteen microsatellite *loci* were detected using fluorescence-based automated capillary electrophoresis (MegaBace 1000 Fragment Analyzer GE). Whilst the genotyping platform has proved to be very promising, it was realized that many genotyping errors occurred; therefore, it could compromise data and invalidate studies for this population. Thus, the allelic typing underwent cautious validation. Null alleles were detected in 10 *loci*, accounting for 32.7% of the total alleles. Ten *loci* did not segregate in Mendelian ratios. Although some procedures have been developed in order to restrain genotyping errors, it was verified that for the multiplex system of genotyping applied, as well as for the volume of samples involved, none of them were effective. Here, it is reported the protocol for correcting the main causes of errors in an independently manner, and also the results of the genotyping that will be used for a next genetic structure research.

Conclusion

Fluorescence-based automated capillary electrophoresis proved to be a fast and efficient method for genotyping microsatellites on a large scale sample, provided that careful analysis of allelic typing is done in order to prevent frequent errors found in multiplex *loci* system.

Keyword: *Pinus taeda*; microsatellite; null allele; genotyping errors; population genetics

Background

Microsatellites have been increasingly popular in genetic mapping and population genetic analysis due their typical high degree of polymorphism and genomic abundance. However, the development and use of the microsatellite markers remain a challenge in many species for many reasons, and genotyping errors are one of the main problems.

Genotyping error occurs when the observed genotype of an individual does not correspond to the true genotype. [1] is an example of the first scientists to note that laboratory errors could result in mismatches in *pedigree* data. An error can be indicated if the experimental genotype is incompatible with reliable independent evidence, such as pedigree data [2].

Null allele of microsatellites is mainly caused by the non-amplification of one allele due to the mutations at PCR priming site. Null alleles are generally inferred at population level by the occurrence of non-amplifying individuals or a surplus of homozygotes at a given *locus* [3]. The inheritance of null alleles can be verified by segregation analysis without ambiguity using full-sib families [4], but this is not the case for half-sib.

Measurements on DNA tissue extracted from Antarctic fur seals [5], as well as from brown bears [6], detected an error rate of up to 0.8% per microsatellite *locus*. In 1997, a genotyping study revealed a striking new model for chimpanzee mating behavior by indicating that half the offspring of a community were sired by males from outside the group [7]. It soon turned out that this conclusion [8] resulted from genotyping errors which led to an erroneous paternity exclusion. Publications involving plant species and genotyping errors were not found.

These are just examples of some serious effects such errors can have on important biological issues. However, few studies in population genetics and evolution quantify the rate of genotyping error [6,5] that would ensure the reliability of the inferred biological conclusion. Moreover, there is no consensual strategy or strict standard for limiting or quantifying the occurrence of main types of error [9]. In [2] published paper, the authors examined the causes and consequences of genotyping errors and have recommended to limit their occurrence and their effect on the resulting biological message.

In this study, a MegaBace 1000 Fragment Analyzer (GE) was used to detect and read the amplified fragments of 15 microsatellite *loci*, in 120 open-sib progenies, from a breeding population from a private program of southern Brazil. A multiplex protocol was elaborated to genotype 3 groups with 5 *loci* each, in a sample of 1,130 individuals of *Pinus taeda* L.

To ensure the reliability of the inferred genotypes, this study main objective, besides detecting the microsatellite polymorphism, is to validate the trustworthy of the polymorphism characterized for the studied population resultant of the allele typing procedure.

Results and Discussion

Calibrating the Microsatellite Multiplex System Protocol

To calibrate the microsatellite multiplex protocol, preliminary tests were performed with the view to verify that amplification occurred at each fluorescence *locus* separately. After best PCR at termocycler conditions were determined for each *locus*, it was elaborated various combinations in order to check which groups of markers could be amplified together, without causing any overlap in fragment size. Before testing the protocol on the platform MegaBace, an agarose gel was run to determine which could be the most promising *locus* combinations.

To have a previous verification of the allele calling performance for each *locus* by the Fragment Analyzer software from MegaBace 1000; and also, to observe how it segregates in a pilot sample, it was used available samples of a controlled cross or a full-sib progeny. The full-sib progeny was represented by 5 individuals, plus their maternal ancestor (the paternal genotype was not informed). This maternal genotype was one of the 120 mothers of the open-sib progenies. This test was firstly performed in MegaBace by genotyping each *locus* separately (see Figure 1), and then, the same genotypes were evaluated using the multiplex system established.

To determine the number of null alleles per *locus* and the frequency of segregating within the population, we used the software MLTR version 3.4. Nevertheless, it was noticed that the results were not as good as if we do it manually, by applying the method of [10]. When segregation was detected by the significance of G test within the progeny, suspected homozygotes were corrected for a null allele as "0 ". Among the 15 *loci*, null alleles were detected for 10 *loci*, which were 66.6% of the *loci* and 32.7% of the total alleles, respectively. On the Hardy-Weinberg analysis, all *loci* virtually showed significance in light of the X^2 test, within each progeny, indicating an excess of homozygotes, even after some cases of null alleles were fixed.

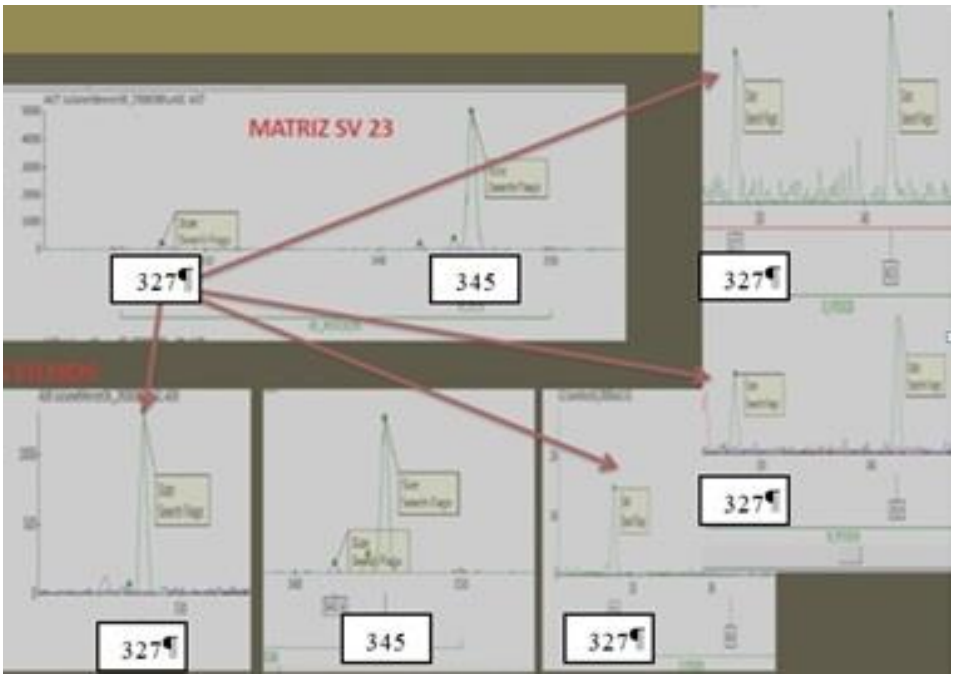


Figure 1. An example of the independent *locus* allele segregation analysis. The images show the genotypes detected in the Fragment Analyzer (GE) for one maternal ancestral (larger image) and its five full-sib progeny. (*Locus* 01 was considered in this image).

After manually correcting the null alleles by estimating the maternal ancestral and testing the alleles segregation within families, this dataset was retested again. Using the MLTR version 3.4, the retest was also implemented to determine the number of null alleles per *locus* that were still presented. Results were still indicating the presence of a single null allele in each *locus*, but its presence would not be a problem, since no frequency higher than 0.02% was detected.

Another pilot experiment was implemented at MegaBace 1000 to access the alleles from 23 maternal ancestral of the 120 open-sib progenies, using the same multiplex system protocol established. The results for those mothers were compared to the same genotypes which were manually inferred, and they proved to have a coincidence higher than 93% in average for all *loci*. This made the inference of the 120 maternal ancestral more reliable.

Regenotyping of missing data was executed in MegaBace using other 6 plates (570 cells). Also, it was included 20 samples randomly chosen within the whole population of open-sib in order to determine if the regenotyping was coincident. It was detected an error of 8% in average for the allele calling.

The Chi-square test showed that 10 *loci* (1, 4, 6, 7, 8, 9, 19, 20, 21 and 24) significantly deviate from the Mendelian expected ratio (P<0.05). For *locus* 7, it was detected a duplication confirmed by the analysis of the maternal segregation within the progenies. It could be taken to be two independent *loci*, since no linkage disequilibrium was detected for them. Based on this evidence, they were referred as being *locus* 7a and 7b.

Microsatellites, due their large amount of variability, have the capacity to inform about duplication. Microsatellites are typically genotyped using PCR primers that amplify all segments of a genome that has coincident nucleotides sequence with the primers sequence. When a microsatellite that lies in a duplicated autosomal region of the genome is applied in a diploid individual, four genomic fragments may be amplified [11].

The MegaBace 1000 automatic Fragment Analyzer used in this study provided a precise fragment sizing (almost 0.4 bp – Figure 2). However, in some cases it was too responsive, making the genotyping across samples difficult to analyze. Hence, after 6 consecutive analysis, the fragment Analyzer peakfilter could be refined (Figure 3) which turned this task easier and more accurate.

The results for the descriptive statistics of each *locus* considered in a single population of 1,130 samples were executed in the GDA software (Table 2). All the 15 *loci* were polymorphic, although *loci* 1, 8 and 17 were removed from the analysis due their excess of missing data. In Table 2 the fixation index of *loci* 23 and 24 were excluded, since the result for them was invalidated by their linkage disequilibrium in prior results.

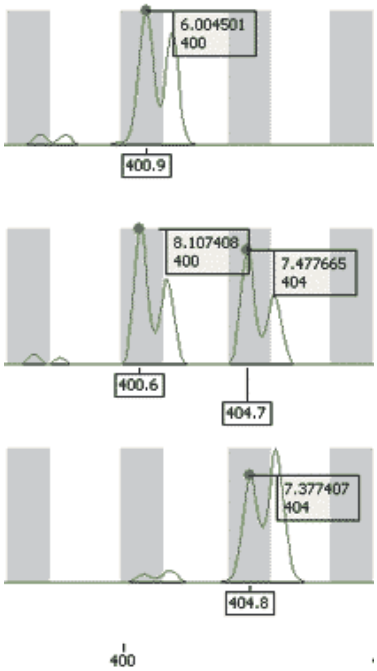


Figure 2. An example of the fragment sizing. The peaks in the image are resultant of the locus 01 genotyping for three random samples in MegaBace 1000 automatic Fragment Analyzer.

Table 1. Descriptive statistic for the microsatellite *locus*. Analyses were run in GDA for the samples of 1,130 individuals trees of the open-sib progenies.

Locus	**%miss**	**N**	**P**	**A**	**Ap**	**He**	**Ho**	**f**
*3*_PtTX3002	27,7	804	1	6	6	0,59	0,52	0,13
*4*_PtTX3088	34,7	729	1	3	3	0,54	0,14	**0,74**
*6*_PtTX3091	27,3	816	1	6	6	0,57	0,24	**0,58**
*7a*_PtTX3098	39,1	697	1	5	5	0,51	0,36	0,30
*7b*_PtTX3098	42,6	649	1	6	6	0,60	0,31	0,49
*9*_PtTX2037	27,0	812	1	6	6	0,64	0,37	0,42
*18*_RPtest5	55,7	499	1	4	4	0,36	0,10	**0,71**
*19*_2n11e	24,6	848	1	5	5	0,61	0,38	0,39
*20*_PRtest8	41,9	657	1	4	4	0,41	0,40	0,01
*21*_PRtest11	61,6	433	1	5	5	0,34	0,18	0,47
*22*_8934M	16,2	947	1	4	4	0,39	0,50	**0,28**

%Miss: % missing data; *N*: the mean sample size over all *loci*; *P*: the proportion of polymorphic *loci*; *A*: the mean number of alleles per *locus*; *Ap*: the mean number of alleles per polymorphic *locus*; *He*: the expected heterozygosity; *Ho*: the observed heterozygosity; and *f*: an estimate of the fixation index using the method of moments.

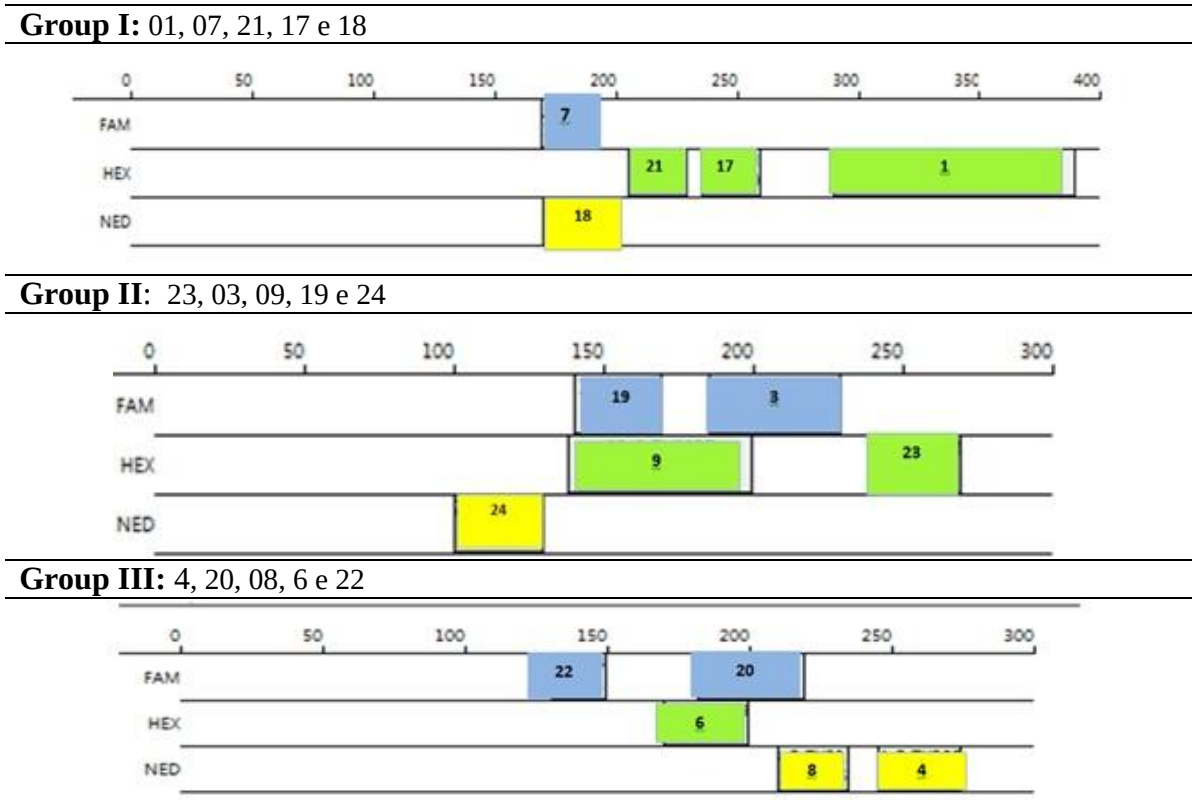


Figure 3. Three groups of multiplex microsatellite *loci* containing 5 *locus* each. The *loci* in each group are represented by their distinct size in base pairs (bp) and their fluorescent label (Fam: blue, Hex: green, Ned: yellow). The images represent the peakfilters of each multiplex system that are available in the Fragment Analyzer (MegaBace 1000).

Causes of genotyping errors

Our own experience during the experiments described above made it possible to observe that genotyping errors result from diverse, complex and sometimes cryptic origins. When an error is detected, the first step is to clearly identify its cause, so that experimental protocols can be improved to reduce error rates. Grouping errors into discrete categories according to their causes is challenging, since different and innumerous causes may interact and generate an error. Some sources of errors may exist due to low quality or quantity of the DNA in the sample, biochemical artifacts, or due to human factors.

Low DNA quantity and/or quality are known as causes that promote genotyping errors. A low number of target DNA molecules in an extract results from either extreme dilution of the DNA or from degradation, which leaves behind only a few intact molecules. Both conditions favor allelic dropouts and false alleles [12]. On average, the quality of the DNA samples detected were good.

The DNA isolation protocol was applied for a large sample size - 1,130 individuals. The protocol procedure allowed isolating DNA with an average of 36 samples per day. Although not statistically tested, it was noticed that the quality of DNA evaluated in agarose gel tends to be influenced by the day it was extracted. This observation leads to the conclusion that, for larger size samples, it may be most convenient to implement a protocol which enables the least human manual activity for the DNA isolation.

Biochemical artifacts

At the end of the elongation step of a PCR, the *Taq* DNA polymerase tends to add a non-template nucleotide (usually an adenine) to the 3′end of the newly synthesized strand [13]. This '+A artifact' is common, and creates an artifactual band or peak on the readout gel or trace, respectively. The relative proportions of the true fragment and the +A artifact fragment are very sensitive to the sequence of the 5′end of the primer used in the genotyping assay, but also to PCR conditions and to the long elongation periods that promote the +A artifact. In such a context, this biochemical artifact represents an important cause of genotyping error. Figure 5 presents these extra peaks detected in the genotyping results.

Human error

Unexpectedly, in the few studies developed to analyze the precise causes of genotyping error, the main cause was related to human factors. In their impressive study on microsatellite genotyping errors used in paternal exclusion in the Antarctic fur seal [5] attributed 80.0%, 10.7%, and 6.7% of the errors to scoring, data input and allelic dropouts, in corresponding order. The remaining 2.7% probably resulted from sample mix-up, pipetting error or contamination.

However, scoring errors might also be an important issue in the automated and semi-automated scoring of fluorescence profiles [2]. For example, human subjectivity during manual scoring represented the main source of discrepancy between the AFLP data sets that were generated by independent scorers which were using the same electropherograms [6].

A number of six alleles scoring were executed for the whole breeding population considering all analysis mentioned before. Observing the results in table 2, it can be noticed that the number of alleles have changed with those validations, which means that the error in allele calling was of 37.34%, since the mean allele number in the initial alleles scoring was 8.3 and after validation it was reduced to 5.2.

This means that researcher's expertise and standards have an influence over the selection of the true allele. Allele calling has also been identified as a potential problem in SNP studies [14]. Therefore, amid various causes of error, allele calling might be the most important to emphasize. Obviously, the risk of human scoring error strongly depends on the quality of data. To access this error, four independent evaluators carried out genotyping, and the results for the alleles calling was observed for all *loci*. The difference in the number of alleles was huge. Also, there was a tendency to detect a much larger number of alleles for the less experienced appraisers.

Table 2. The number of alleles per *locus* that was characterized in the initial and in the final allele typing analysis.

Code	*Locus*	Initial Alleles Number	Final Alleles Number
Locus 1	PtTX3026	10	7
Locus 3	PtTX3002	13	6
Locus 4	PtTX3088	9	3
Locus 6	PtTX3091	8	6
Locus 7[a]	PtTX3098	12	5
Locus 7[b]	PtTX3098	7	6
Locus 8	PtTX3118	8	5
Locus 9	PtTX2037	7	6
Locus 17	PtTX3025	9	8
Locus 18	RPtest5	7	4
Locus 19	2n11e	7	5
Locus 20	PRtest8	6	4
Locus 21	PRtest11	8	5
Locus 22	8934M	6	4
Locus 23	9317M	8	5
Locus 24	PRtest1	7	4
Mean value		8,3	5,2

Conclusions

This study shows that microsatellites are good markers for family analysis in the breeding population of *Pinus taeda* open-sib progenies, despite of the occurrence of null-alleles, segregation distortions or genotyping errors.

The fluorescence labeling and the automatic genetic analyzers provided an efficient method for genotype multiple microsatellite *locus* in a single run, though firstly, the multiplex system must be well calibrated, and the DNA must be of good quality. Although the system detection is recognized as being automated, the sensitivity for PCR artifacts negatively influences the reading of fragments, resulting in high frequencies of false alleles. In the present study, it was concluded that the main source of error is allele calling by human experience.

Even with a high level of null alleles detected, they could be corrected by the family segregation analysis, so their presence did not affect the usefulness of the corresponding *locus*.

Methods

Reference family and DNA Isolation

The references families used in this study were 120 twelve-years-old open-sib progenies, installed in a complete block design, with 9.4 plants per family. Each progeny was originated in one of five forestry plantations (FP) in Santa Catarina state.

The genetic material used as a way of control during the multiplex system protocol establishment, was consisted of one full-sib progeny of five individuals originated from controlled pollination, although, in this study, only the maternal sample was already known. Besides this maternal sample, other 22 maternal tree samples were available. These 23 maternal trees were the same maternal ancestors of the 23 progenies within the whole breeding population of 120 open-sib progenies. The sample material was frozen needles tissue, kept in a -20ºC freezer.

DNA was isolated using the CTAB protocol of [15] modified by BITTENCOURT, J.V.M. The needles were pulverized and digested in 2×CTAB buffer (100 mM Tris-HCl pH 8.0, 1,4M NaCl, 20 mM EDTA, 2% CTAB) at 65ºC for 30 minutes. Solution was extracted with chloroform and then precipitated using isopropanol. Pellets were washed in 96% ethanol, dried and suspended in 50 µl TE (1M Tris-HCL Ph 8,0, 0,5M EDTA). Extracted DNA was stored at -20ºC.

Microsatellite amplification

The microsatellite primers used and its reference literature are summarized on Table 3. PCR reagents were purchased from QIAGEM (Uniscience), and primers were customized from RWGenes. One of the two primers in every *locus* is 5': Hex, Fam or Ned-labeled (Table 3).

It was used the QIAGEN Multiplex PCR Kit (product code 206143) for individual and multiplex microsatellite *loci* PCR, containing: 2x QIAGEN Multiplex PCR Master Mix and RNase-free water. The 2x QIAGEN Multiplex PCR Master Mix contains: HotStarTaq® DNA Polymerase, dNTP Mix and PCR Buffer. Information that the 2x QIAGEN Multiplex PCR Master Mix final concentration contained 3 mM $MgCl_2$ is available. The QIAGEN® Multiplex PCR Handbook can be downloaded at <www.qiagen.com>.

Table 3. Fluorescent microsatellite *locus* descriptions. Before the underline: the number used to reference each *locus* during this research; after: the name of the *locus;* and in brackets the number of Genbank accession; authors who described the loci in publication; the fluorescent label, the melting temperature (Tm) and the multiplex system number adopted in the study.

Locus Name and Number adopted	Label	Tm	Multiplex
*1*_PtTX3026 (AF143971)	Hex	58^0C	I
*9*_PtTX2037 (AF143959)	Hex	58^0C	II
*17*_PtTX3025 (AF143970)	Hex	60^0C	I
*3*_PtTX3002 (AF277846)	Fam	58^0C	II
*6*_PtTX3091 (AF277848)	Hex	58^0C	III
*4*_PtTX3088 (AF277843)	Ned	58^0C	III
*7*_PtTX3098 (AF277847)	Fam	58^0C	I
*8*_PtTX3118 (AF277845)	Ned	58^0C	III
*18*_RPtest05 (BV728798)	Ned	60^0C	I
*20*_RPtest08 (BV728799)	Fam	60^0C	III
*21*_RPtest11 (BV728796)	Fam	60^0C	I
*24*_RPtest01 (BV728795)	Ned	60^0C	II
*19*_2n11e (AA556153)	Fam	60^0C	II
*22*_8934M (AA739818)	Fam	60^0C	III
*23*_9317M (AA740072)	Hex	60^0C	II

Genotyping

PCR products were separated and detected using a MegaBace 1000 (GE Healthcare) with the Fragment Analyzer Software of its platform. The capillary was filled with MegaBace Long-Read Matrix, a denaturing polymer. The standard size was MegaBace ET550-ROX.

Appropriate volume of PCR products were diluted in ddH2P 20x with 0,01% twin, 0,22 μl of ET-ROX, with a final volume of 8 μl per sample, that was finally denatured at 94ºC for 3 minutes and chilled on ice for 3 minutes. Each genotyped plate had 96 cells with 95 samples and 1 control containing all the reagents described, except for the PCR product. The plate was finally loaded onto MegaBace 1000, with injection time of 80 seconds and 3Kvolts, and a run time of 120 minutes with 8Kvolts. The amplified fragments were then detected and collected by the MegaBace 1000 software. The number of genotyped plates was 36 for the breeding population, plus 6 plates which were used in the regenotyping procedure.

Segregation and null allele identification

Microsatellite *loci* were scored as codominant markers. Different alleles (fragments) of each *locus* were named according to the size in base pairs. Once the null allele was

identified and confirmed, it would be denoted by 0, otherwise, as lost allele -9. The segregation of microsatellite *loci* were tested by firstly inferring the maternal ancestral. Although, before that, it was assured that the 23 maternal genotypes used as control matched the inferred ones. Then, all the 120 maternal genotypes were generated. Mendelian proportions (1:1, 1:1:1:1, or 1:2:1) of the progeny genotypes for *loci* were examined using Chi-square test at $p \leq 0.05$. Linkage and Hardy-Weinberg disequilibrium analysis were carried out using GDA 1.1 (Lewis and Zaykin, download at: http://hydrodictyon.eeb.uconn.edu/people/plewis/software.php). To determine the number of null alleles per *locus* and the frequency of segregating in the population, we used the software MLTR version 3.4 [21].

Competing interests

The authors declare that they have no competing interests. The population samples used in the study are legal property of a Brazilian private program, and it was previously integrated in a major forestry breeding funded project, coordinated by HIGA, A. R.; doubly supported partly by private initiative, partly by the funding agency FINEP, executed at LAMEF (Laboratory of Forest Genetics and Breeding, Department of Forestry - UFPR).

Authors' contributions

JRM performed all experimental work, analyses, and also wrote the paper. MdeLLA contributed with some of the data analysis and provided advice. All authors read and approved the final manuscript.

Acknowledgements

We thank the tree samples facility provided by the private company program. We thank Dr. Alexandre Magno Sebben (Instituto Florestal de São Paulo) and Dr. Gustavo Avelar (GE HealthCare) for valuable insights and contributions during the experiments. We also thank the students from LAMEF lab: Nicole Munhoz, Adriane Partala and Thais Vagaes for helping with the DNA isolation. This work was supported by FINEP and Fundação Araucária.

References

1. Thompson EA. **A paradox of genealogical inference**. *Adv Appl Probab* 1976, **8**:648–-650.

2. Pompanon F, Bonin A, Bellemain E, Taberlet P. **Genotyping errors: causes, consequences and solutions**. *Nature Reviews* 2005, **6**:847-859.

3. Primmer CR, Moller AP, Ellegren H. **Resolving genetic relationships with microsatellite markers: a parentage testing system for the swallow *Hirundo rustic*.** *Mol Ecol* 1995, **4**:493-498.

4. Li L, Guo X, Zhang G. **Inheritance of 15 microsatellites in the Pacific oyster *Crassostrea gigas*: segregation and null allele identification for linkage analysis**. Chinese Journal of Oceanology and *Limnology* 2009, **27**:74-79.

5. Hoffman JI, Amos W. **Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion**. *Mol Ecol* 2005, **14**:599–612.

6. Bonin A, Bellemain, E, Bronken Eidesen, P, *et al*. **How to track and assess genotyping errors in population genetics studies**. *Mol Ecol* 2004, **13**:3261–3273.

7. Gagneux P, Woodruff DS, Boesch, C. **Furtive mating in female chimpanzees**. *Nature* 1997a, **387**:358–359.

8. Gagneux P, Boesch C, Woodruff DS. **Microsatellite scoring errors associated with noninvasive genotyping based on nuclear DNA amplified from shed hair**. *Mol.Ecol.* 1997b, **6:** 861–-868.

9. Ewen KR, Bahlo M, Treloar SA, *et al*. **Identification and analysis of error types in high-throughput genotyping**. *Am J Hum Genet* 2000, **67**:727–736.

10. Gillet E, Hattemer HH. **Genetic analysis of isoenzyme phenotypes using single tree progenies**. *Heredity* 1989, **63**:135-141.

11. Zhang K, Rosenberg NA. **On the genealogy of a duplicated microsatellite**. *Genetics* 2007, **177**:2109–2122.

12. Taberlet P, Griffin S, Goossens B, *et al*. **Reliable genotyping of samples with very low DNA quantities using PCR**. *Nucleic Acids Res* 1996, **24:**3189–3194.

13. Magnuson VL, Ally DS, Nylund SJ, Karanjawala ZE, *et al*. **Substrate nucleotide-determined non-templates addition of adenine by Taq DNA polymerase: implications for PCR-based genotyping and cloning**. *BioTechniques* 1996, **21**:700–709.

14. Ghosh S, Karanjawala ZE, Hauser ER, *et al*. **Methods for precise sizing, automated binning of alleles, and reduction of error rates in large-scale genotyping using fluorescently labeled dinucleotide markers**. *Genome Res* 1997, **7**:165–178.

15. Doyle JJ, Doyle, JL. **Isolation of plant DNA from fresh tissue**. *Focus* 1990, **12**:13-15.

16. Elsik CG, Minihan VT, Hall SE, Scarpa A.M, Williams CG. **Low-copy microsatellite markers for *Pinus taeda* L**. *Genome* 2000, **43**:550–555.

17. Zhou Y, Gwaze DP, Reyes-Valdés MH, Bui T, Williams CG. **No clustering for linkage map based on low-copy and undermethylated microsatellites**. *Genome* 2003, **46**:809–816.

18. Kutil BL, Williams CG. **Triplet-repeat microsatellites shared among hard and soft pines**. *J Hered* 2001, **92**:327-332.

19. Chagné D, Echt C, Richardson T, Plomion C, *et al.*: **Cross-species transferability and mapping of genomic and cDNA SSRs in pines**. *Journal Theor Appl Genet* 2004, **109**:1204-1214.

20. Allona I, Quinn M, Sederoff R, Whetten RW, *et al.* **Analysis of xylem formation in pine by cDNA sequencing**. *Proc Natl Acad Sci* 1998, **95**:9693-9698.

21. Ritland K. **Extensions of models for the estimation of mating systems using *n* independent *loci***. *Heredity* 2002, **88**:221-228.

CAPÍTULO II

ESTRUTURA GENÉTICA DE UMA POPULAÇÃO SELECIONADA DE *Pinus taeda* L.

MERCER[1], Juliane Rezende; LIMA[2], Milena de Luna Alves; ALMEIDA[3], Marina Isabel Mateus; HIGA[4], Antonio Rioyei; GLIENKE, Chirlei[5]

A estrutura genética de uma população selecionada de *Pinus taeda* L. brasileira, representada por 120 progênies de polinização aberta foi determinada usando inferência bayesiana em análises de agrupamento multicluster, com 15 *loci* microssatélites (SSR) em 1130 amostras, assumindo equilíbrio de Hardy-Weinberg (Structure 2.3) [1], e assumindo endogamia (InStruct) [2]. Os cones das sementes que originaram a população-alvo haviam sido coletados de 120 antepassados maternos, as quais haviam sido selecionadas fenotipicamente para características de volume de madeira em 5 diferentes florestas plantadas (FPs) no sul do país, cerca de 15 anos atrás. Houve também uma seleção para produtividade de madeira na população selecionada, resultante de um teste de progênies, e foram selecionadas a melhor (*i*) e a (*ii*) segunda melhor árvore por repetição dentro de progênie. Para certificar os resultados obtidos, foi adotado um procedimento de *learning-samples*, buscando encontrar o número de grupos genéticos (ou K) mais provável. Somente então, os parâmetros genéticos foram inferidos. A primeira hipótese refutada foi a de que o valor mais provável K=5, obtido pelas estatísticas *ad hoc* de PRITCHARD [3] e EVANNO [4], coincidia com as cinco FPs - uma vez que as FPs *a priori* foram assumidas como sendo de 5 origens ou procedências distintas. Foi aproveitada a estrutura familiar da população para inferir os genótipos dos ancestrais maternos. Concluiu-se que a geração materna é mais provável de ter sido originada pela mistura de 3 origens distintas de sementes. Que existam 5 grupos genéticos (K=5) na população de progênies, e que estes tenham se formado pela ocorrência de cruzamentos preferenciais, e, também por uma forte pressão de seleção dentro de progênies. As árvores de melhor valor genético (*i*) mantiveram uma maior variabilidade genética em comparação ao das árvores de segundo melhor desempenho (*ii*); com valores superiores em heterozigosidade observada e em número de alelos maternos mantidos. O modelo de migração que melhor explica os resultados é o de zona-de-contato. Os coeficientes de endogamia de Wright (F_{ST}) foram 2 a 3 vezes maiores nas progênies do que em relação aos da geração materna. A relevância dos resultados obtidos e a forma com que foram explorados são de interesse tanto para estudos voltados à Genética de Populações, quanto para os programas de melhoramento. Uma vez que podem ser uma ferramenta para auxiliar no monitoramento da variabilidade genética da população selecionada.

PALAVRAS-CHAVE: *Pinus taeda* L., população selecionada, estrutura genética, microssatélites, parâmetros genéticos, inferência bayesiana.

Abstract*

GENETIC STRUCTURE OF A *Pinus taeda* L. BREEDING POPULATION

MERCER[1], Juliane Rezende; LIMA[2], Milena de Luna Alves; ALMEIDA[3], Marina Isabel Mateus; HIGA[4], Antonio Rioyei; GLIENKE, Chirlei[5]

The genetic structure of a Brazilian breeding population of *Pinus taeda* L., represented by 120 open-sib progenies was determined using bayesian inference in multicluster analysis with 15 microsatellite *loci* (SSR) in 1130 samples, assuming Hardy-Weinberg (Structure 2.3) [1], and assuming inbreeding (InStruct)[2]. The seed cones that gave rise to the target population had been collected from 120 maternal ancestors, which had been phenotypically selected for wood volume in five different forestry plantations (FPs) in the south of Brazil, about 15 years ago. There was also a selection to improve wood volume production within the breeding population, resultant from a previous progeny test, so it was selected the first best (*i*) and second best (*ii*) tree per block within each open-sib progeny. To make the results reliable, it was adopted a procedure of "learning-samples ", in order to find the number of genetic clusters or most likely K. Then, the genetic parameters were inferred. The first hypothesis that was rejected was that the most probable value of K= 5, obtained from *ad hoc* statistics of PRITCHARD [3] and EVANNO [4], was coincident with the five FPs - since the FPs were, *a priori*, assumed to be from 5 different backgrounds or origins. It was used the familiar structure of the population to infer the genotypes of maternal ancestors. It was concluded that the maternal generation is the most likely to have been planted by the mixture of three different seed sources or origins, that there are five genetic groups (K= 5) in the population of progeny, and that they have been formed from the occurrence of assortative mating, and also, from a strong pressure in the selection within families. The trees with the best genetic value (*i*) maintained a higher genetic variability when compared to the trees of second best performance (*ii*), with higher values of heterozygosity and of numbers of maternal alleles that were kept the same. The migration model that best explains the results is the contact zone model. The inbreeding coefficients of Wright (F_{ST}) were 2-3 times higher in offspring than in relation to the maternal generation. The relevancy of the results and the way they were explored may be of value both for studies of Population Genetics, as for plant breeding programs, since they help monitoring the population's genetic variability during generations of selection.

KEYWORDS: *Pinus taeda* L., breeding population, genetic structure, microsatellites, genetic parameters, bayesian inference.

*/ O artigo está sendo preparado a ser submetido para publicação na revista "Genetica", a qual publica artigos que evolvem os temas: genética, dinâmica populacional e evolução; incluindo estudos de estrutura genética de populações, evolução do genoma, especiação, comportamento, conservação. Seja qual for o táxon considerado. Endereço Eletrônico: *http://www.springer.com/life+sciences/journal/10709?detailsPage=editorialBoard#*

Introduction

Pinus taeda Linnaeus is a monoecious conifer, diploid organism with a high rate of cross-fertilization. It is originated from the Southern and Southeastern United States, and it was first introduced in Brazil in the 40s. Its wood characteristics are desirable for industry panels and sawn timber. It is a fast-growing species and it has being well adapted to temperate climate, what makes it a strong candidate for the establishment of planted forests in Southern Brazil. From a total of 5.74 million hectares (ha) of existing forest plantations in Brazil in 2006, the area with *P. taeda* in the southern states were 1,520,000 hectares [5]. Most of these plantations were established with genetically improved seeds, produced in clonal orchards – orchards established by breeding programs, based on phenotypic selection of trees, which explored the genetic variability among and within populations planted in the form of commercial plantations.

Evolutionary forces acts on population dynamics promoting genetic signatures in allele frequencies that are transmitted to succeeding generations. Mainly resulting from the action of artificial selection, one of the most drastic form of expression of these forces are manifested in the genetic and phenotypic populations of domesticated species. Molecular tools and Biometrics can characterize all, or part of such signatures, generating knowledge that can greatly contribute to the efficiency of genetic improvement of the breeding programs, through monitoring the genetic variability of the population during the generation of selection.

"Population of Genetics" studies the variation in allele frequencies between and within populations, being "Wright's F statistic" the most common method used to infer genetic structure in populations. F_{ST} denotes the Wright inbreeding coefficient [6]. To calculate these indices, we must first define what are the genetic groups or subpopulations, and thus, use the individuals' genotypes in the inference of allele frequencies of these groups. However, the estimation of F_{ST} can be susceptible to discrepancies and distortions, thus an essential prerequisite to any inference about the genetic structure of populations is the definition of the populations themselves. The cluster analyses are usually used for this purpose.

There are two broad types of clustering methods. One is based on the analysis of genetic distance and the other is based on model of Hardy-Weinberg. In the method of distance, algorithms are used to construct phylogenetic trees [7]. In bayesian clustering methods, many have been developed [8;9]. The methods implemented by STRUCTURE software are one of the most recent and most promising in literature [10;11;12].

The algorithm of this bayesian model assumes the existence of k's population (k may be unknown), and each is characterized by a set of allelic frequencies per *locus*. Individuals in the sample are probabilistically assigned to populations or jointly to two or more populations if their genotypes indicate they are hybrids or "*admixtures*" [3], although the inference of the actual number of clusters K's in an existing set of data is notoriously difficult and computationally challenging. However, the Bayesian paradigm represents an alternative of what should be done, since it is possible to assign an *a priori* probability distribution for K [Pr (K | X)] and use them to infer the most likely number of K due to the distribution of likelihood *a posteriori* [Pr (X | K)]. However, this posterior distribution might be particularly dependent on the modeling assumptions adopted.

Evanno *et al.* [4] used various dispersal scenarios from data and generated a model based on individuals migration. It was found that in most cases the estimates of Pr (X | K) do not provide a correct estimate of the real number of clusters, K. The authors then, propose an *ad hoc* statistic (K) based on the rate of change in log of probabilities reported for successive values of k's, *a priori*. In this study, it was implemented the *ad hoc* statistical of EVANNO [4], and also, an amendment to it to determine the most probable number of K. It was noticed that the small change promoted a better fit for the results of STRUCTURE for this population. The pattern of population migration resulted from the analysis leads to a stepping-stone or contact zone model, with at least two dimensions.

Assortative mating induces correlations in allelic states within and between *loci* that can be explored to understand the genetic structure of natural populations [13]. For many species, it is considerable interesting to quantify the contribution of two forms of non-random mating in order to understand the patterns of genetic variation under inbreeding (mating among relatives), and under fragmented population (limited dispersal of gametes).

GAO *et al.* [2] extended the popular bayesian clustering approach of STRUCTURE [3] to simultaneous inference of inbreeding or selfing rates and classification of the source population using *multilocus* genetic markers. This is achieved when the Hardy-Weinberg equilibrium is no longer assumed within groups and then, the expected genotype frequencies are estimated based on rates of inbreeding or selfing. Researchers demonstrated the need for such procedure when it was indicated that selfing leads to spurious signals of structured population using the standard STRUCTURE algorithm, with a tendency to bias toward spurious signals of admixture.

The subdivision of the population can generate a linkage disequilibrium between distant chromosomal regions. Also, as long as it concerns to association studies, this can lead

to spurious associations between traits and polymorphisms, unless the population structure is properly characterized in the statistical analysis. Plant populations are often characterized in genetic structures as hard to immediate resolution, mainly because mating among relatives and hybridization with different origins may occur [14]. GAO *et al.* [2] measured the performance of the proposed method using extensive coalescent simulations and demonstrated that the approach can correct this bias.

The objective of this work was to study the genetic structure of a population that consists of open-sib progenies of *P. taeda*, using microsatellite markers (SSR: Simple Sequence Repeats) using a bayesian approach that, assumes Hardy-Weinberg equilibrium, to determine the groups or sub-populations that are most real. Then these results will be contrasted with bayesian analysis assuming inbreeding to verify if the genetic groups remain similar.

Methods

Reference Family and Microsatellite locus Genotyping

The reference family used in this study consisted of 120 open-sib progenies at 12 years old, installed in a complete block design with five repetitions. The total number of samples were 1,130 trees. Each progeny was originated within one of five forestry plantations (FPs) in Santa Catarina, a Brazilian southern state, about 15 years ago (Table 1). Geographic localization is not available for FPs.

The seed cones that gave rise to the target population had been collected from 120 maternal ancestors, which had been phenotypically selected for wood volume at the five FPs.

Table 1. The individuals' trees distribution of the open-sib progenies, which represent the breeding population, in relation to their Forestry Plantation (FP) origin.

FP	N_{os}	N_{mp}	N_{tp}	% size
FP_1	28	9,5	266	23,5
FP_2	18	9,27	167	14,8
FP_3	28	9,42	264	23,4
FP_4	5	9,2	46	4,1
FP_5	41	9,44	387	34,2
Total	120	9,42	1130	100%

Number of open-Sib progenies (N_{os}), mean size of progeny in number of individuals (N_{mp}), total number of individuals within FP (N_{tp}) and the percentage of each FP individuals' origin to the total breeding population.

There was also a genetic selection for wood volume improvement in order to form the breeding population, which was previously determined by the private program progeny test. The selection defined the first best (*i*) and second best (*ii*) tree within each open-sib progeny, in each of the five blocks of a complete block design. The genetic values determined by the breeding program for the performance results of the progeny test are unknown.

The sample material was frozen needles tissue, kept in a −20ºC freezer. DNA was isolated using CTAB. A microsatellite multiplex system protocol was used for genotyping in MegaBace 1000 (GE Healthcare). The microsatellites *loci* used with their respective number of accession in Genbank within parenthesis, were: ***PtTX3026*** (AF143971); ***PtTX3002*** (AF277846); ***PtTX3088*** (AF277843); ***PtTX3091*** (AF277848); ***PtTX3098*** (AF277847); ***PtTX3118*** (AF277845); ***PtTX2037*** (AF143959); ***PtTX3025*** (AF143970); ***RPtest5*** (BV728798); ***2n11e*** (AA556153); ***PRtest8*** (BV728799); ***PRtest11*** (BV728796); ***8934M*** (AA739818); ***9317M*** (AA740072) e **PRtest1** (BV728795).

Genetics and Multicluster Bayesian Analyses

Firstly, the segregation of microsatellite *loci* was tested by inferring the maternal ancestral within each open-sib progeny. Linkage and Hardy-Weinberg disequilibrium analysis were carried out using GDA 1.1 (LEWIS and ZAYKIN, download at: http://hydrodictyon.eeb.uconn.edu/people/plewis/software.php). The structure was characterized using bayesian inference in multicluster analysis with 15 microsatellite *loci* (SSR) in 1130 samples, under two criteria: assuming Hardy-Weinberg (software STRUCTURE 2.3 – 1. Available at: <http://pritch.bsd.uchicago.edu/structure.html>), and assuming inbreeding (software INSTRUCT - 2. Available at: <http://cbsuapps.tc.cornell.edu/InStruct.aspx>). The permutation analysis for testing the individual's clusters assignment were implemented in CLUMPP 1.1.2, a Cluster Matching and Permutation Program [15]. Two *ad hoc* statistics were used: the *ad hoc* statistic of PRITCHARD [3], and the *ad hoc* statistic of EVANNO [4].

To determine the number of burn-in and the number of repetitions in the *ad hoc* statistic of PRITCHARD it was made a pilot study with 20 independent runs in Structure. It was observed that the posterior distributions respectively converge with 10,000 and 50,000.

In order to test the *ad hoc* statistical method of EVANNO [4], 20 independent runs for each K*i a priori* were performed, being *i*: 1 to 21, using the admixture model and correlated allele frequencies of STRUCTURE. A pilot study was conducted with the view to determine

the number of burn-in and repetitions, which were respectively 10,000 and 50,000, used in the 420 independent analyses.

Before implementing the EVANNO *ad hoc* statistic, an exploratory study was conducted to determine the nature of dispersion for the values of the logarithms of likelihood represented by the posterior probabilities [Pr (X|K)]. It was concluded that the mean was not a good estimator for the dispersion. Thus, a change in the authors' equations was implemented, which basically consisted in using the median values instead of the mean. Therefore, instead of using the ΔK introduced by the authors, it was referred here as Δk_m.

The equations used by Structure to infer the F_{ST} or the inbreeding coefficient of Wright, which is a measure of correlations between populations, the He or the expected heterozigosity, the coancestry coefficients (q) and the genetic distance using the allelic frequencies are fully described in [16;3]. The equations and the inference of genetic parameters of Instruct are fully described in [2].

In order to determine and perform permutations that verify which of the arrays indicated the best individuals for each of the five clusters, and also to determine the *admixtures*, it was used the Clumpp 1.1.2 software with 10,000 permutations for the 30 Q matrices generated in 30 races the STRUCTURE 2.2.

After considering the initial analysis, the mode 5 of the INSTRUCT software proved to have the best performance for this population. This mode has estimated rates of selfing or inbreeding using back-reflection. A number of 100 analyses were performed to estimate the parameters of inbreeding and clustering that provides the best probabilities of likelihood for the posterior K. The number of MCMC iterations was 50,000, 10,000 of burn-in, and 10 for the thinning period.

Hypothesis Statements for the Learning Sample Methodology

i. If Forestry Plantations represent different origins, then five coincident genetic groups are expected in the breeding population.

ii. If only the *admixtures* individuals from the breeding population are analyzed, then it is more straightforward to determine the most genetical distinct Forestry Plantations.

iii. If the *ad hoc* Statistic of Pritchard is efficient in determine the most likely number of genetic groups, then the most real K will be determined.

iv. If the *ad hoc* Statistic of Evanno is efficient in determine the most likely number of genetic groups, then the most real K and the migration pattern will be determined for the breeding population.

v. If the modification in the *ad hoc* Statistic of Evanno is more efficient in determine the most likely number of genetic groups, then the most real K and the migration pattern will be determined for the breeding population.

vi. If the *ad hoc* Statistic of: Pritchard, Evanno and Evanno's modification are coincident, then the most real K result is consistent.

vii. If the genetic groups of the maternal origin are determined, then it is possible to verify how it behaved in the progenies generation.

viii. If it is observed a tendency in distinct patterns of maternal genetic groups to hybridizations in the *admixtures* individuals of the breeding population, then an indication of preferential crosses within maternal genetic groups may have existed.

ix. If the criterion of Hardy-Weinberg equilibrium assumed in the bayesian analyzes causes a statistical bias toward the number of genetic groups inferences, then assuming the inbreeding assumption will result in a distinct number of K and in a different distribution of individuals within these K's.

x. If the number of microsatellite *loci* is not enough to determine the genetic structure of the global breeding population, the modified *ad hoc* statistic of Evanno will oscillate the most real K results in the analyses with different number of *locus*.

xi. If sample size influences the stabilization in the bayesian analyses posterior parameters inferences distribution of probabilities, then it is possible to determine for which sample size the number of *loci* will be enough for the genetic parameters estimation.

xii. If all above hypothesis are tested then it will be possible to determine the most real K to characterize the genetic structure of the breeding population, and also will be possible to give a biological explanation for the fragments formation.

Results and Discussions

Forests Plantations (FP's) represent distinct genetic groups?

The STRUCTURE analysis started by using the admixture model with correlated allele frequencies, with no information about the population. The objective was to test the hypothesis that FP's are distinct genetic groups. The results with highest posterior

probabilities indicated the existence of 5 groups or K=5 in 100 consecutive analyses. Next, analyses were performed by informing the FP origin and the results did not converge among K= 5 with 5 FP's (Figure 1). Figure 2 presents the distances in allelic frequency of these analyses.

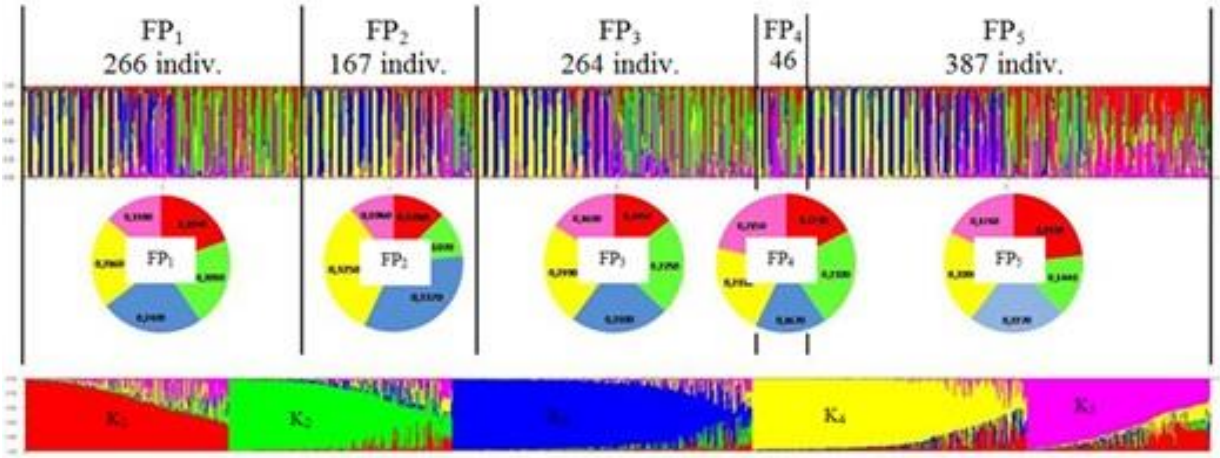


Figure 1. Distribution of the 5 genetic clusters (K's) in relation to each of the five forestry plantations (FPs). Upper graph: 5k's a function of FP's; lower graph: proportion of coancestry (Q) of the 5k's in 1130 samples. Pie charts: the contribution of each k within FP.

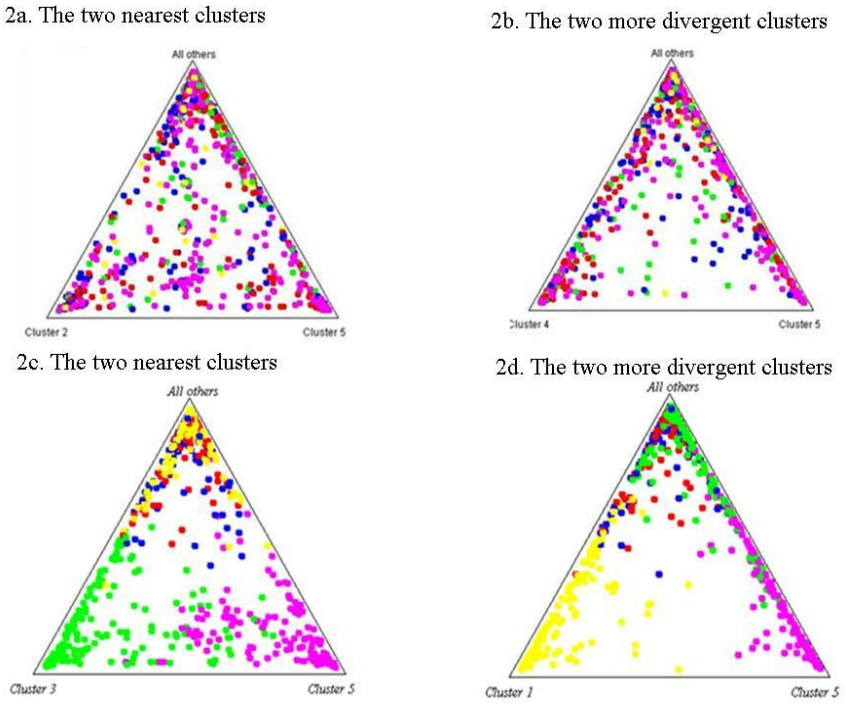


Figure 2. The distances in allelic frequency from the genetic clustering analyses of the breeding population. In the triangles above (2a and 2b) the five colors distinguish the forestry plantations (FP's); bellow, (2c and 2d), the five colors represent the most likely 5K's genetic groups. Each point in the triangle expresses the distance of the individual to the genetic clusters (vertices).

These results indicate that the FPs are a mosaic of these five genetic clusters, and that the proportion of contribution of each cluster in the FPs is what sets them apart. In the results in which the analyses were run, with the FPs origin informed, there is little difference in genetic diversity between the FP's, for both numbers of polymorphic alleles - an average of 4.55 alleles per *locus* and for the expected and observed heterozygosity, which is very unlikely to happen considering that the FPs were from distinct origins. The fixation index was lower for FP_2, being the most divergent among the FPs (Table 2).

To verify the gene flow between the FPs, the analyses considered the FP origin information, but it was settle as not informed for the individuals classified as being admixture in the first analysis. The results estimated the confidence interval of the admixture individuals coancestry to each of the five genetic clusters, and also the probabilities for the number of past generations in which the hybridization event have happened (parents, grandparents and great-grandparents). The results indicated that 60% of *admixtures* were generated in the parental generation during the maternal ancestral pollination in the forestry plantations.

To see how the *admixtures* behaved, all the nom-admixture individuals resulted from previous analysis, and also the smallest forestry plantation population (FP_4), were excluded from this analysis. The best posterior probability for this dataset resulted for K=3 (Figure 3). The most distinct forestry population was FP_2 and $FP_{3;}$ the majority of FP_5 being a hybrid between FP_2 and FP_3. Although half of the individuals in FP_1 have the same hybrids, this population have a distinct genetic group, which also appears in the other four forestry plantations and almost did not hybridizes among the other genetic groups. However, deeper analyses in order to test this hypothesis of gene flow could not be stated since no information was provided to the geographical locations of the forestry plantations.

Table 2. Descriptive statistics for the admixture individuals detected for the breeding population separated by the five Forestry Plantations (FPs).

Population	N	n	n %	P	A	Ap	He	Ho	f
FP_1	266	163,46	61,5	1,00	4,62	4,62	0,49	0,33	0,32
FP_2	167	108,69	65,1	1,00	4,92	4,92	0,51	0,38	0,26
FP_3	264	173,54	65,7	1,00	4,54	4,54	0,51	0,35	0,31
FP_4	46	29,38	63,9	1,00	3,92	3,92	0,51	0,34	0,34
FP_5	387	245,62	63,5	1,00	4,77	4,77	0,48	0,31	0,35
Mean	*226*	*144,14*	63,8	*1,00*	*4,55*	*4,55*	*0,50*	*0,34*	*0,32*
TOTAL	1130	720,69							

(N): number of individuals in each FP; (n) number of admixture, or individuals that had less than 95% of coancestry to a determined cluster, (n%) the percentual value of n; (P) the proportion of polymorphic loci; (A) the mean number of alleles per locus; (Ap) the mean number of alleles per polymorphic locus; (He) the expected heterozygosity (He); (Ho) the observed heterozygosity; (f) and an estimate of the fixation index using the method of moments in GDA.

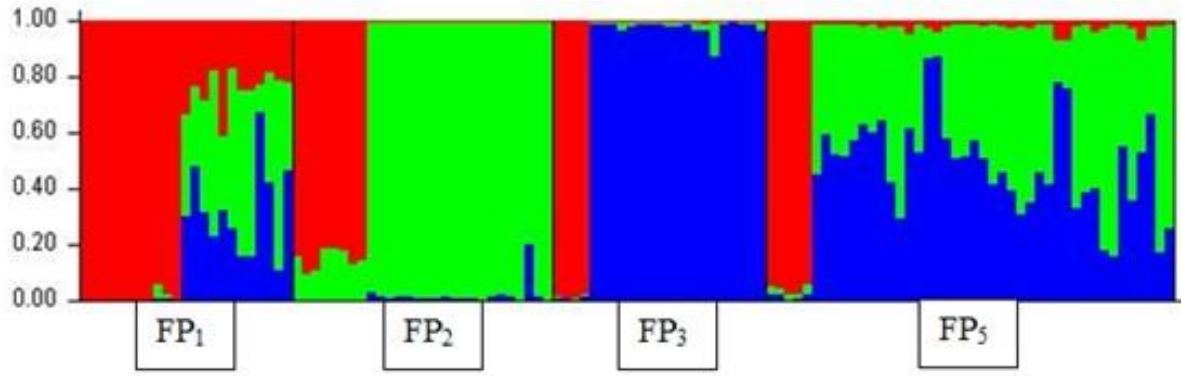


Figure 3. Analyses for the *admixtures* detected in previous analysis for K=5. It was excluded the samples of non-*admixtures* and data from FP_4. For the present analysis, the best posterior probability was K=3.

Ad hoc Statistic of Pritchard

To perform the procedure in the statistical *ad hoc* of PRITCHARD [3], it was executed 10 runs for each K *a priori*, and these K's were from 1 to 20. The values of the average for those posterior probabilities for each K [Pr (X|K)] are presented in Table 3. The best [Pr (X|K)] was detected for K=19 (-12,238.4), but the most appropriate value, according to the literature, would be the smallest number of cluster that captured most of the genetic structure; after that, it occurs an increasing pattern for the next K's [Pr (X|K)] until a plateau is reached. Hence, that happens for K=5 (-13337.3). The adjustments of alpha and F_{ST} values during the burn-in MCMC, can be observed in Figure 4, that resulted to the best posterior distribution analysis of K_1, K_5 and K_{19}. Note that for K= 5, they showed the best adjustments.

Table 3. The posterior probability [Pr (X|K)] on average for 10 consecutive analyses of each K *a priori*. Being K's from 1 to 20. Also resumes the Wright's inbreeding coefficients (FST) and inferred values of alpha.

K	ln Pr(X|K)	Alpha	F_{st}_1	F_{st}_2	F_{st}_3	F_{st}_4	F_{st}_5	F_{st}_6
1	-16672.1	-	0.1080					
2	**-15056.9***	**0.0666**	**0.1620**	**0.1878**				
3	-14416.8	0.0473	0.192	0.169	0.295			
4	**-13541.4**	**0.0427**	**0.3256**	**0.2054**	**0.3522**	**0.2757**		
5	**-13337.3**	**0.0415**	**0.3186**	**0.3885**	**0.3017**	**0.3778**	**0.1964**	
6	-13084.8	0.0418	0.3037	0.1942	0.3892	0.3861	0.3701	0.3786
7	-12814.6	0.0402	0.4003	0.4197	0.3179	0.4532	0.3536	0.3944
8	**-12700.5**	**0.0395**	**0.2619**	**0.2847**	**0.4235**	**0.5584**	**0.4340**	**0.4157**
9	-12553.3	0.0380	0.5705	0.3574	0.4223	0.4284	0.3811	0.5048
10	-12937.5	0.0384	0.4098	0.3624	0.5221	0.4361	0.4866	0.4091
11	-12469.0	0.0374	0.4400	0.3308	0.4720	0.4553	0.5905	0.4737
15	-12319.5	0.0367	0.5479	0.4228	0.4794	0.5518	0.5449	0.3863
19	**-12238.4**	**0.0372**	**0.3775**	**0.5804**	**0.5108**	**0.4279**	**0.4106**	**0.5439**
20	-12693.2	0.0374	0.4783	0.5684	0.5109	0.4521	0.4646	0.5389

**/In bolt numbers are the best posterior values.*

Inference for the most real K using ad hoc statistic of Evanno

EVANNO *et al.* [4] proposed a statistic to determine the most probable number of K that could be graphically interpreted. In the author's publication, they have simulated different populations' sizes, with different types and numbers of markers for three types of migration patterns: islands, islands and hierarchical contact-zone (Stepping-Stone).

In figure 5, the graph depicts the results for ΔK_m , which demonstrates that the highest peak is for K=2. In figure 6, there is the same graphic with K_2 not included, in order to have a better visualization of the second highest peaks detected for K_4, K_{19} and K_8. The graphic shows two analyses. In the first one, it was used all 15 *loci* (red: squares); in the second analysis, it was considered the duplication at *locus* 07 as being two independent *loci* (7a and 7b), and also it was excluded the *loci*: 01, 08 and 17 – for the higher levels of missing data – totalizing 13 *loci* in the second tests of ΔK_m (blue triangles). It was concluded that both analyses – using 15 or 13 *loci* – provide similar results.

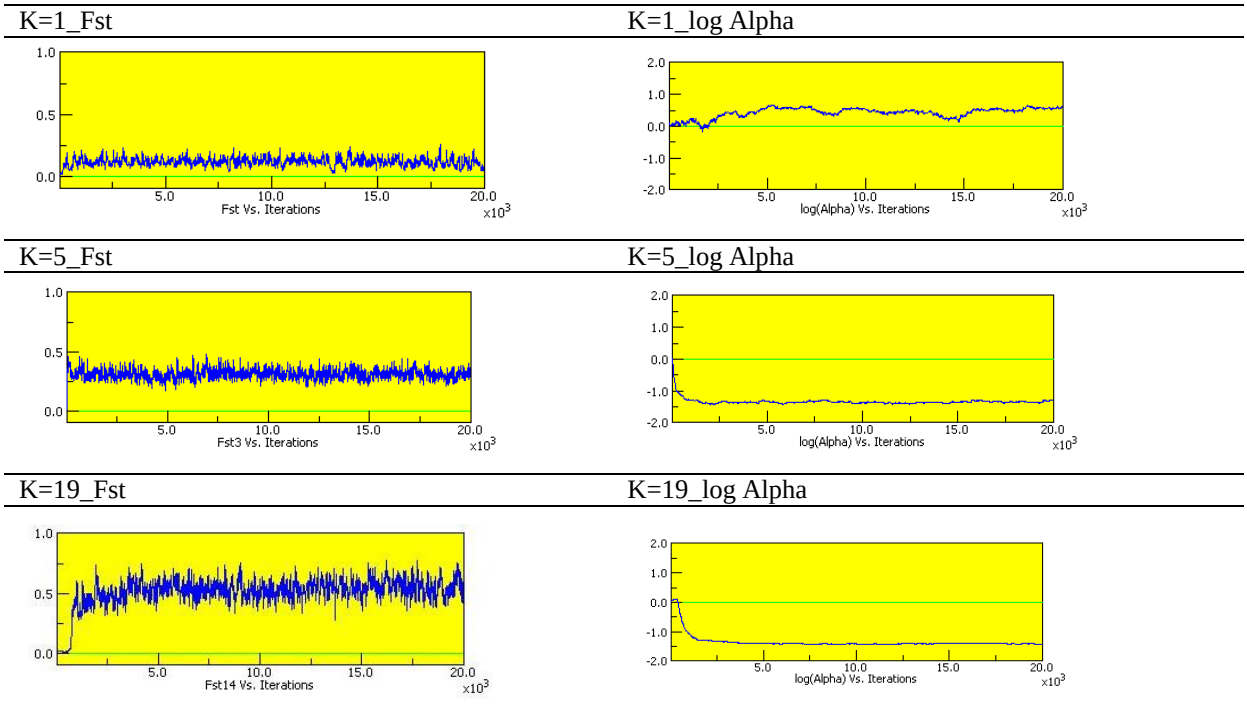


Figure 4 Graphics resulted for log alpha and F_{ST} distribution during the burn-in period of the K_1, K_5 and K_{19}.

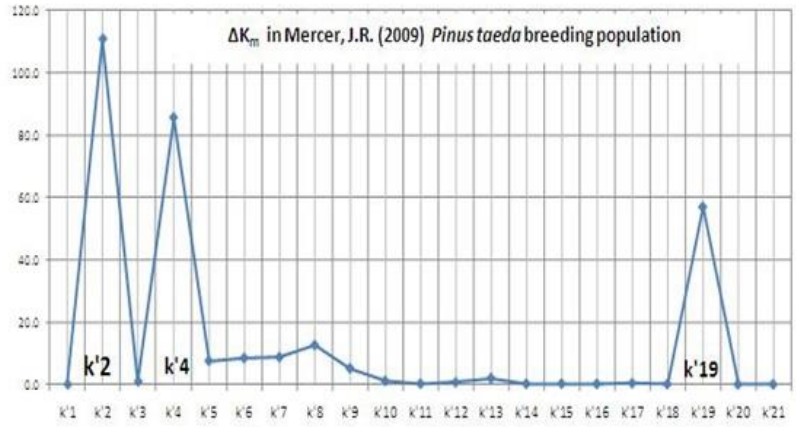


Figure 5. The graph shows the results for ΔK_m with the highest peak detected for for K_2.

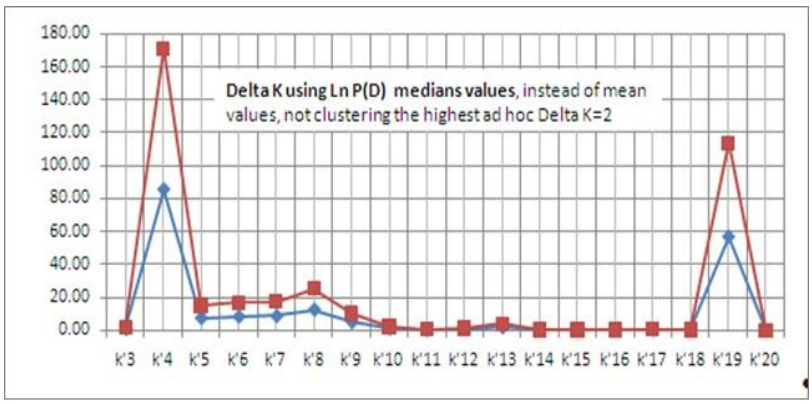


Figure 6. ΔK_m using medians values. Not including K_2 (the strongest signal), but including the second higher peaks at K_4, K_{19} and K_8, in analyses of ΔK_m using 15 *loci* (red: squares) and 13 *loci* (blue triangles).

The greatest magnitude of ΔK_m at K_2 indicated a migration pattern similar to a stepping stone or contact zone model, as shown in [4], separating two groups of four populations with even a second dimension of structure within these 4 populations that may reach 19 subpopulations.

Then, it was tested how the individuals that were first assigned for the K=5, from the *ad hoc* statistic of PRITCHARD, behave in the partition of populations and subpopulations detected by the ΔK_m. After countless analyses, the following conclusions were obtained and the genetic parameter of this analysis is demonstrated on Table 4.

The greatest peak in magnitude ΔK_m for K_2 was remarkably originated by a division in the selection that had occurred within progenies (see 4th line on table 4 for K_2). This means that, even without informing the first best trees *i* and the second best trees *ii* from the progeny

test, they clearly subdivide the breeding population into two greatly differentiated genetic groups or fragments.

Some relevant aspects can be determined by the observation of the division of distribution for the trees *i* and *ii* among the five clusters in which they represent almost the total number of individuals within these groups (see 2nd line on ~~at~~ table 4 for K_5). The results showed that individuals from the *ii* trees are much more structured and have less genetic diversity than the subset tree *i*. On average, the F_{ST} for trees *i* was 0,24 and He 0,45; for trees *ii* the F_{ST} and He were 0,36 and 0,4, respectively.

Figure 7 summarizes the results and adds the genetic distances among all substructures detected. It makes it more visually comprehensible. These results came from the analysis that leads to the stepping stone model, expressing the distance in allele frequency divergence among all the 19 substructures.

Though the results from the two *ad hoc* statistics seem different, after performing an analysis within groups and subgroups, the conclusion is that both of them lead to the same results, and that K=5 is the most real number of genetic groups to represent the breeding population.

Genetic structure of the maternal generation

The inbreeding coefficient "F_{ST}" for the five clusters of the breeding population ranged from 0,19 to 0,42 indicating a very differentiated population, with low pollen flow within it. For the genetic structure detected in the maternal generation it was found that the most real number of genetic groups is K=3, and a genetic inbreeding coefficient ranging 0,01 to 0,15 (Table 5).

Also, analyses were conducted in order to verify the relations between the maternal and its descendent~~s~~ genetic clusters. It was indicated a tendency of assortative mating occurrence among specific genetic clusters that originated some open-sib progenies, although these results are not shown in the present document.

Table 4. Analysis within each of the genetic structure and substructures detected using the *ad hoc* statistic of EVANNO *et al.* (2005), in relation to the partioning of the K=5.

Posterior Ks[1]	F_{ST} and He within subcluster										Total
	***k'$_1$**		**k'$_2$**		**k'$_3$**		**k'$_4$**		**k'$_5$**		
	F_{ST}	He	**F_{ST}**	He	**F_{ST}**	He	**F_{ST}**	He	**F_{ST}**	He	
5	**0.2**	0.5	**0.42**	0.4	**0.32**	0.4	**0.2**	0.44	**0.3**	0.4	1.130
	158 (*i*)[2]		240 (*ii*)		251 (*i*)		195 (*i*)		286 (*ii*)		
4	*k'$_3$*+ *k'$_4$*[3]		**0.40**	0.4	**0.25**	0.4	**0.2**	0.5	**0.28**	0.41	1.130
3	*k'$_4$*		**0.36**	0.4	*k'$_4$*		**0.19**	0.5	**0.28**	0.41	1.130
2	*k'$_4$*		*k'$_5$*		*k'$_4$*		**0.2** Trees *i*	0.5	**0.2** Trees *ii*	0.5	1.130

**/ k' represents the genetic groups (K) previously detected using the ad hoc statistic of PRITCHARD. [1]/ The distribution of individuals considering distinct number of populations, with their respectively values of F_{ST} and He. [2]/ i and ii are the best and second best trees, and their assingment distribution within k'.[3]/ the indication of the individuals new k' assingment with the reduction of K populations. If it appears more than one k' ,it indicates that the previous k' have split to the indicated k's.*

Table 5. F_{ST} values and expected Heterozygosity (He) between the three maternal clusters and the five genetic clusters of the breeding population.

F_{ST1}	He_1	F_{ST2}	He_2	F_{ST3}	He_3	F_{ST4}	He_4	F_{ST5}	He_5
120 mothers (K=3) [*-3184.9*[1] */ 242.7*[2] */ 0.188*]									
0.01	0.5	0.13	0.53	0.16	0.45				
1130 OP indv. (K=5) [*-13263.4*[1] */1120.2*[2] */0.046*[3]]									
0.19	0.5	0.42	0.38	0.32	0.38	0.24	0.44	0.30	0.41

[1]/ LnP(D) posterior probability of prior k [2]/ Var[LnP(D)] of the probability. [3]/alpha value estimative.

Bayesian Analysis assuming inbreeding

The best posterior probability in the analyses of INSTRUCT software or assuming inbreeding was, again, obtained for K= 5, with the lower value for the log likelihood in mean and in variance (mean and variance -6855.3: 13710.5)

The result of the inbreeding bayesian analysis that had the best convergence for the Markov chain is being considered. This is determined by the value of the Gelman-Rubin (GR) Statistic. This statistic measures the convergence of the log-likelihood. The value for this chain was 1.022.

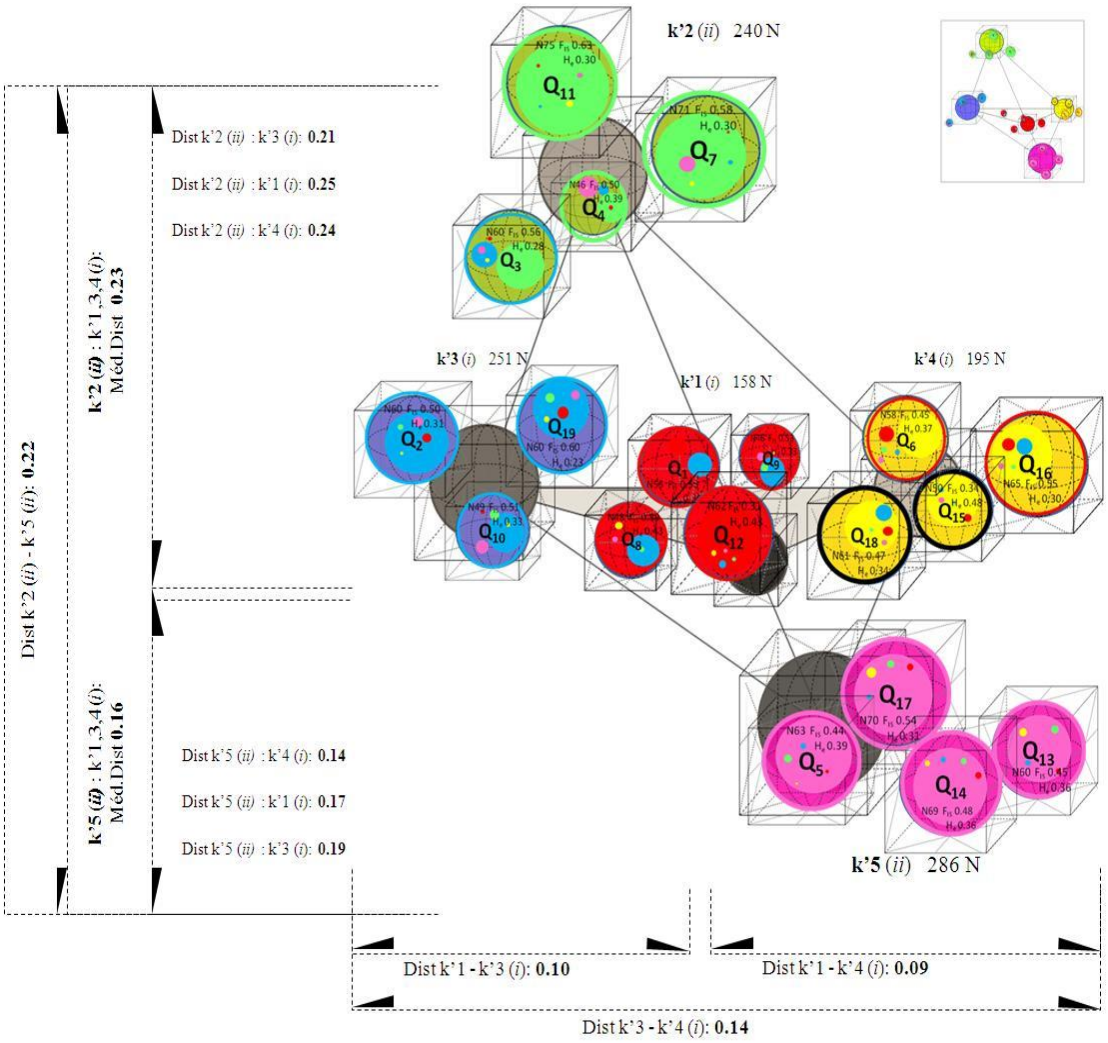


Figure 7. Migration model proposed based on the distribution of the five genetic groups (K=5) and how is it distributed within 19 subpopulations

Figure 8 presents the five most likely resulting clusters detected by INSTRUCT for the coefficients of F_{ST}, in order of increasing inbreeding, considering the contribution of each of the five K's or subpopulations from STRUCTURE and CLUMMP (K=5).

Table 6 summarizes the obtained results, depicting the percentage of contribution from each of the five genetic groups obtained in STRUCTURE (K=5); the five clusters assuming inbreeding obtained with INSTRUCT.

It is noticed that the group with the lower inbreeding coefficient – cluster 1 of INSTRUCT with a F_{ST} of 0.022 (Figure 8), consisted of 56% of individuals which belonged to K_4 resulting from STRUCTURE, and the K_4 are almost exclusively formed by the first best genetic trees (*i*) – see table 4. Regarding the group of higher inbreeding coefficient, or the cluster 5 of the INSTRUCT – F_{ST} of 0.227, it was formed by the contribution of 72% of the

individuals in the K_2 of STRUCTURE, and the K_2 is almost exclusively formed by the second best genetic trees (*ii*). Once more, it is relevant to emphasize that the analyses were conducted without prior knowledge of the trees *i* or *ii* classifications.

Some studies report that individual heterozygosity, at apparent neutral microsatellite markers, is correlated with key components of individual fitness such as fecundity [19], disease resistance [20] and lifetime reproductive success [21]. The verification that the higher heterozygosity was detected for *i* trees in comparison with the group of *ii* trees – on average – is surprising, since wood volume is not considered an adaptive trait. On the other hand, the group with the *ii* trees had the higher inbreeding coefficient (F_{ST}) – on average –, what could bring the idea that the poor wood volume performance comparing with *i* trees may be due inbreeding depression, as the underlying mechanism. When a genome carries deleterious recessive alleles, individuals with above average homozygosity will be relatively unfit, and this is the basis of classical inbreeding depression or genome-wide heterosis [22].

The distribution of the five resulting clusters using INSTRUCT was also very similar to the STRUCTURE, indicating again that forestry plantations are a mosaic of all five groups that were coincident in analysis assuming Hardy-Weinberg (STRUCTURE) and assuming inbreeding (INSTRUCT).

STRUCTURE generates clusters based on both transient Hardy–Weinberg disequilibrium (HWD) and linkage disequilibrium (LD) caused by admixture between populations. The program works by clustering individuals in groups in which both linkage and HWD are minimized and, as a result, the presence of LD in the data improves clustering results [16]. On the other hand, 'strong' LD or departure from Hardy–Weinberg equilibrium

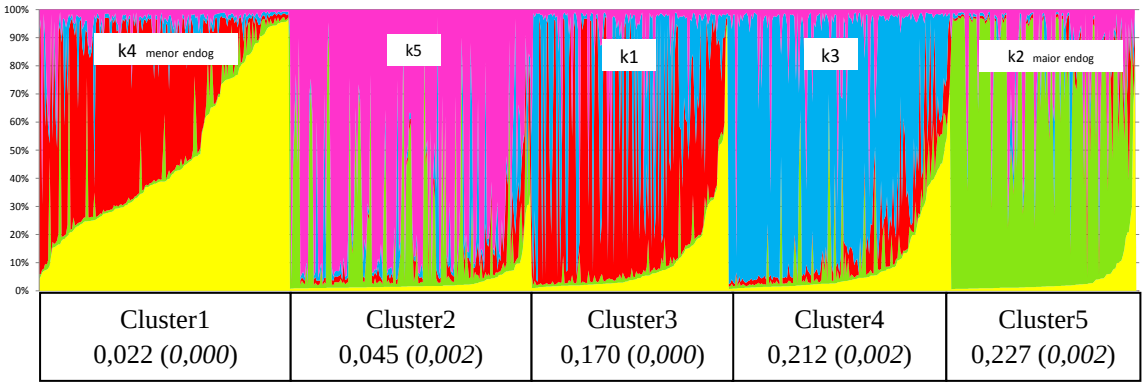


Figure 8. Bar graph of the five resulting clusters in the bayesian analysis assuming inbreeding (INSTRUCT). It demonstrates the 5 clusters coinciding with the K= 5 obtained on the analysis in STRUCTURE software. Below, each cluster is presenting the results of F_{ST} and variance for each cluster.

could lead to an overestimation of the number of detected clusters [16]. Since almost all microsatellite *locus* of this study departure from Hardy-Weinberg equilibrium, it was expected that those K=5 were overestimated and, if so, INSTRUCT would detect a lower number of genetic clusters. However, it did not occur. In fact, the results for number of most probable K were similar in both methods.

In the analysis made by STRUCTURE, the inbreeding coefficient varied from 0,19 to 0,42 for 1,130 individuals and 15 microsatellite *loci*. In the analysis that assumes this INSTRUCT inbreeding, the F_{ST} values ranged from 0,022 to 0,227, however, these results were analyzed with only 738 individuals (that have the lower number of missing data for the genotyped alleles) from the existing 1,130, and with a reduction of 15 to 10 in the number of microsatellite *loci* (by excluding the *loci*: 01, 06, 08, 17 and 21). This has occurred because the stability of the convergence for the log of likelihood of the posterior distributions of K occurred for this dataset configurations in the INSTRUCT analysis. Table 7 summarizes this result. The F_{ST} values assuming inbreeding were lower than assuming Hardy-Weinberg, but the convergence for these analyses could only be reached and exclude individuals with higher number of missing data.

Table 6. Proportion of the individuals' contribution in the bayesian analyses under both criteria. c' express in percentage to form each genetic cluster in the analysis assuming inbreeding (INSTRUCT, 1st column and 1st line) in, and assuming Hardy-Weinberg equilibrium (STRUCTURE, 2nd column), and the number of total individuals (N) that were used in this analysis.

c'	k'5	c'1	c'2	c'3	c'4	c'5	N
1	4	0,30	0,06	**0,53**	0,05	0,06	171
2	5	0,05	**0,65**	0,03	0,08	0,18	162
3	1	**0,51***	0,08	0,11	0,27	0,03	132
4	3	0,09	0,07	0,09	**0,68**	0,07	149
5	2	0,04	0,13	0,04	0,05	**0,73**	124

**The bolt numbers denotes the higher genetic group contribution in number of individuals.*

Tabela 7. Comparing Fst values of the open-sib progenies assuming Hardy-Weinberg equilibrium (STRUCTURE, 2nd line) and assuming inbreeding (INSTRUCT, 3rd line) with and their maternal genotypes inferred (STRUCTURE, 1st line).

Independent Structure analyses	k	LnP(D)	Var [LnP(D)]	Fst_1	Fst_2	Fst_3	Fst_4	Fst_5
120 mothers	3	-3184	242.7	0.099	0.126	0.156		
OSF 1130 indv.	5	-13353	1108.0	0.375	0.326	0.384	0.194	0.290
OSF 738 indv.	5	-6905,2	13810.4	0.170	0.227	0.212	0.022	0.045

Thus, this result confirms that the population is structured into 5 groups that were coincident in both criteria, under Hardy-Weinberg and under inbreeding. These results confirm that the reduction in expected heterozygosity is probably caused by the occurrence of higher rates of selfing and assortative mating.

In a publication of [17] as well as in other references [18], F_{ST} values from 0,01 to 0,05 would be expected for populations of *pinaceaes* species of open cross pollinating under natural conditions; and from 0,05 to 0,30 for divergent populations. These values, compared to the results of the analysis, indicate the high degree of differentiation obtained for this population of breeding.

Pinus taeda has historically large, interconnected populations which extend along the US Atlantic seaboard, from Maryland to Florida, and westward to central Texas. [23] tested the allelic diversity in samples from natural stand selections that represent populations prior to intensive plantation establishment and domestication, and defined the existence of five genetic populations with a high level of gene flow within populations. Although the five Ks detected in the present study cannot be compared to the five populations detected in [23], such coincidence draws attention.

Conclusions

The bayesian method of clustering analysis using the 15 microsatellite *loci* led to a consistent global knowledge of the genetic structure of the population.

Based on the observed results, it was possible to conclude that the five forestry plantations, where the seed sources that gave rise for the studied breeding population was collected (the maternal generation), were a mosaic of 3 genetic groups; and after one generation of pollination at the FPs, the differentiation of fragments became even higher, leading to the five genetic groups determined at the breeding population.

Both the bayesian analysis that assumed Hardy-Weinberg equilibrium and the one that assumed inbreeding resulted in five coincident genetic groups, being the most likely K to represent the breeding population.

The F_{ST} values that assumed inbreeding were lower than the ones that assumed Hardy-Weinberg equilibrium. Hence, the definition for the genetic parameters of F_{ST} and H_e will be considered from the Hardy-Weinberg equilibrium analyses, which is: F_{ST}: 0,194 to 0,384 and H_e: 0,38 to 0,5.

It is probable that the highly fragmented breeding population phenomenon detected in the bayesian analyses has firstly occurred due to the selection pressure within open-sib progeny when the program selected the first and second best trees for wood volume, but also, due to the ocurrence of assortative mating phenomenon.

Competing interests

The authors declare that they have no competing interests. The population samples used in the study are legal property of a Brazilian private program, and were previously integrated in a major forestry breeding funded project, coordinated by HIGA, A. R.; doubly supported partly by private initiative, partly by the funding agency FINEP, executed at LAMEF (Laboratory of Forest Genetics and Breeding, Department of Forestry - UFPR).

Authors' contributions

JRM conceived and directed this study. Also, performed all experimental work, biometric analysis, hypothesis statements and wrote the paper. MdeLLA provided advice. All authors read and approved the final manuscript.

Acknowledgements

The tree samples facility provided by the private company program. The authors gratefully acknowledge the advices of Dr.Juliana Vitória Messias Bittencourt. We thank Dr. Echt Craig (USDA Forestry) and Dr. Kostya Krokovisky (Texas University) for valuable insights for the analysis. This work was supported by FINEP and Fundação Araucária.

References

1. Pritchard, J.K.; Wen, X.; Falush, D. Documentation for structure software: Version 2.3. Software from http://pritch.bsd.uchicago.edu/structure.html. Feb., 2010.

2. Gao, H.; Williamson, S.; Bustamante,C.D. A Markov Chain Monte Carlo Approach for Joint Inference of Population Structure and Inbreeding Rates From *Multilocus* Genotype Data. *Genetics*: **176**, 1635–1651. 2007.

3. Pritchard, J.K; Stephens, M.; Donnelly, P.J. Inference of population structure using *multilocus* genotype data. *Genetics*: **155**, 945-959. 2000.

4. Evanno, G.; Regnaut, S.; Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*: **10**, 1365-1375. 2005.

5. Sociedade Brasileira De Silvicultura. **Fatos e Números do Brasil Florestal**. São Paulo, s.d. Dezembro, 2007. 110p.

6. Wright, S. The genetical structure of populations. *Ann. Eugenics*: **15**, p. 323-354. 1951.

7. Falush, D.; Stephens, M.; Pritchard, J.K. Inference of population structure using *multilocus* genotype data: dominant markers and null alleles. *Molecular Ecology*: **?**, p.895-908. 2007.

8. Banks, M.A.; Eichert, W. Whichrun Version 3.2: a computer program for population assignment of individuals based on *multilocus* genotype data. *J. Heredity*: **91**, p.87-89. 2000.

9. Corander, J.; Waldmannb, P.; Sillanpää, M.J. Bayesian Analysis of Genetic Differentiation Between Populations. *Genetics*: **163**, p. 367-374. 2003.

10. Rosenberg, N.A. *et al.* Empirical evaluation of genetic clustering methods using *multilocus* genotypes from 20 chicken breeds. *Genetics*: **132**, p. 699-713. 2001.

11. Thornsberry, J.M.; Goodman, M.M.; Doebley, J. *et al.* Dwarf8 polymorphisms associate with variation in flowering time. *Nature Genetics*, **28**, p.286–289. 2001.

12. González-Martínez, S.C.; Krutovsky, K.V; Neale, D.B. Forest-tree population genomics and adaptive evolution. *New Phytologist*: **170**, 227-238. 2006.

13. Wright, S. The interpretation of population structure by F-statistics with special regard to system mating. *Evolution*: **19**, 395–420. 1965.

14. Camus-Kulandaivelu, L., Veyrieras, J.B.; Gouesnard, B.. Charcosset, A.; Manicacci, D. Evaluating the reliability of structure outputs in case of relatedness between individuals. *Crop Science*: **47**, 887–892. 2007.

15. Jakobsson, M.; Rosenberg, N.A. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*: **14**, p.1801–1806. 2007.

16. Falush, D.; Stephens, M.; Pritchard, J.K. Inference of population structure using *multilocus* genotype data: linked *loci* and correlated allele frequencies. *Genetics*: **164**, 1567–1587. 2003.

17. Cavalli-Sforza, LL.; Menozzi, P.; Piazza, A. The History and Geography of Human Genes. Princeton University Press, Princeton, NJ. 1994.

18. Yu, J.; Zhang, Z.; Zhu, C. Simulation Appraisal of the Adequacy of Number of Background Markers for Relationship Estimation in Association Mapping. *The Plant Genome*:.**2**, p.63-77. 2009.

19. Amos, W.; Wilmer, J.; Fullard, K. *et al.* The influence of parental relatedness on reproductive success. *Proc. Roy. Soc. Lond.*, **268**, 2021–2027. 2001.

20. Coltman, D.; Pilkington, J.; Smith, J.; Pemberton, J. Parasite mediated selection against inbred Soay sheep in a free-living, island population. *Evolution:* **53**, 1259–1267. 1999.

21. Slate, J.; Kruuk, L.; Marshall, T.; Pemberton, J.; Clutton-Brock, T. Inbreeding depression influences lifetime breeding success in a wild population of red deer (cervus elaphus). *Proc. Roy. Soc. Lond.*: **267**, 1657–1662. 2000.

22. Keller, L.; Waller, D. Inbreeding effects in wild populations. *T. Ecol. Evol.*: **17**, 230–241. 2002.

23. Al-Rabab'ah, M.A.; Williams, C.G. Population dynamics of *Pinus taeda* L. based on nuclear microsatellites. *Forst. Ecol. Manag.*: **163**, 263-271, 28. 2002.

DISCUSSÃO

Os *loci* microssatélites marcados com fluorescência foram caracterizados em eletroforese capilar automatizada (MegaBace 1000 Fragment Analyzer GE), em sitema multiplex contendo cinco primers por grupo, que tiveram suas condições de amplificação em PCR calibradas na presente pesquisa. A tabela 3 (Cap.I) apresenta os fluoróforos de cada *locus*, podendo ser: Hex, Fam or Ned. Contudo, foi observada uma grande quantidade de erros de genotipagem.

Foi considerado erro de genotipagem quando o genótipo conferido a um índivíduo da amostra não corresponde ao verdadeiro genótipo. Neste caso, o genótipo não é verdadeiro, pois foi determinado pela presença de um ou dois alelos falsos. Alelos falsos podem ter ocorrido pela presença de fragmentos amplificados em PCR, que na verdade não eram fragmentos microssatélites, mas sim, artefatos de PCR; como também, pela não amplificação de um ou dois dos alelos, em função da ocorrência de mutação na região do sítio de anelamento do primer microssatélite, os alelos nulos. Além dos erros de genotipagem ocorreram também os alelos perdidos, que representam alelos que não foram amplificados no procedimento de genotipagem por alguma falha que pode ter ocorrido em qualquer uma das atividades envolvidas.

Alguns fatores que podem ter influenciado na ocorrência de erros de genotipagem foram a baixa qualidade e quantidade de DNA, os artefatos bioquímicos e os erros humanos. Porém, por mais que tentativas tenham sido feitas para isolar estes erros no momento da análise dos dados do polimorfismo, percebeu-se que é muito difícil conseguir discriminá-los.

Um fator que pode ter contribuído negativamente no resultado da genotipagem, foi sem dúvida, a qualidade do DNA genômico (TABERLET *et al.*, 1997). Foram observadas muitas amostras de baixa qualidade e, no total, a qualidade não foi padrão para o total da população. Algumas tiveram que ser relavadas com etanol 76%. Por mais que o protocolo de CTAB utilizado (Fig.2.An.2) tenha se mostrado eficiente para um volume pequeno de amostras (Fig.1.An.2), uma recomendação que se faz necessária para volumes maiores é a utilização de protocolo de isolamento de DNA que utilize *kits* comerciais, para os quais normalmente estão envolvidas um menor número de manipulação humana.

O volume médio de amostras de DNA isolado por dia com o protocolo usado foi de 36 amostras. Não foi testado estatisticamente, porém foi percebido que a qualidade de DNA testada em gel de agarose variou de acordo com o dia de realização da atividade. Isso demonstra como a influência da atividade humana interfere no resultado final de isolamento de DNA. Uma imagem do gel de agarose de uma das 12 placas contendo o DNA de toda a população com 95 amostras em cada placa, apresenta uma das imagens utilizada para avaliar visualmente a qualidade e quantidade de DNA, em comparação com o tamanho de base padrão do fago lambda/Hind III (Fig.2.An.2).

Uma fonte que pode promover a ocorrência de erros de genotipagem é o sistema multiplex em si. Se erros de genotipagem são comuns na amplificação de *locus* individuais, estes se acentuam quando na amplificação simultânea de mais de um *locus*. Para calibrar o protocolo multiplex até determinar um adequado, em que os *loci* pouco interferissem entre si na reação de PCR, vários testes foram realizados.

Um fator positivo deste experimento foi a utilização de um *kit* comercial (QIAGEN) para a realização de todas as atividade de PCR, uma vez que todos os reagentes, incluindo a *Taq* polimerase, já estavam sob ótimas condições de concentração em solução. Tal fator, provavelmente, deve ter evitado fontes de variação nos resultados finais de genotipagem.

Inicialmente a amplificação em PCR foi testada para cada *locus* com primers fluorescentes. Após determinada as melhores condições em termociclador para cada *locus*, testes foram realizados para verificar quais poderiam ser amplificados conjuntamente, sem alterar os fragmentos detectados. Assim, inúmeras combinações de *locus* foram testadas, levando em consideração que os fragmentos de *locus* diferentes não poderiam se sobrepor. Porém, se houvesse sobreposição em tamanho de fragmentos, tais *loci* deveriam ter diferentes cores fluorescentes.

Contudo, por mais que esforços tenham sido efetuados para obter o melhor protocolo de calibração que resultaram em três grupos multiplex, fragmentos amplificados em PCR que não correspondem aos verdadeiros fragmentos ainda ocorreram para a genotipagem em larga escala.

A detecção dos fragmentos contidos nas amostras que migram na eletroforese capilar do sequenciador é feita pelo próprio software da plataforma MegaBace 1000, o Fragment Analyzer. Os fragmentos contidos nas amostras podem ocorrer por inúmeras causas e o

desafio é detectar quais são os verdadeiros dos falsos. Os fragmentos são detectados pelo espectro de luminescência que emitem ao refletir o lazer que recebem durante sua migração em eletroforese, e estes migram em diferentes ve*loci*dades de acordo com o tamanho em pares de bases (*bp*). Como os marcadores microssatélites são marcadores codominantes, um indivíduo diplóide pode apresentar dois alelos diferentes, e cada alelo (fragmento amplificado) é determinado e distinguido pelo seu tamanho em pares de bases.

Para verificar como ocorria a detecção de alelos no Fragment Analyzer, foi utilizada uma amostra piloto, representada por um progenitor materno e seus cinco filhos, frutos de um cruzamento controlado por uma única fonte de pólen – progenitor masculino (amostra não disponível). Essa amostra piloto foi genotipada e, para cada indivíduo foi realizado o procedimento de PCR para cada *locus*, para todos os 15 *loci*. A figura 2 (Cap.I) mostra um exemplo de análise para verificar como os alelos segregavam dentro desta família. Após, a amostra piloto foi genotipada pelo sistema multiplex. Foi verificado um bom resultado, e por isso foi aceito o protocolo do sistema multiplex para prosseguir a genotipagem em larga escala.

Picos extras muito próximos ao verdadeiro alelo microssatélite também ocorreram em certos *loci* estudados. A causa dessa ocorrência pode ser por uma tendência da *Taq DNA polymerase* em incorporar um ou mais nucleotídeos extras - normalmente a adenina - na extremidade 3’ da nova cadeia que é sintetizada durante a etapa final de extensão no ciclo de PCR. Este artefato comum é conhecido por ‘*+A artifact*’, e a sua ocorrência pode ser influenciada pela sequência nucleotídica da extremidade 5’ do primer, pelas condições de PCR, e pela duração em tempo na etapa de extensão do PCR (MAGNUSON *et al.*, 1996). A figura 1 apresenta algumas imagens destes picos extras, que geraram certa dificuldade na definição do alelo verdadeiro em certas amostras durante a tipagem dos alelos, promovendo o descarte destes resultados.

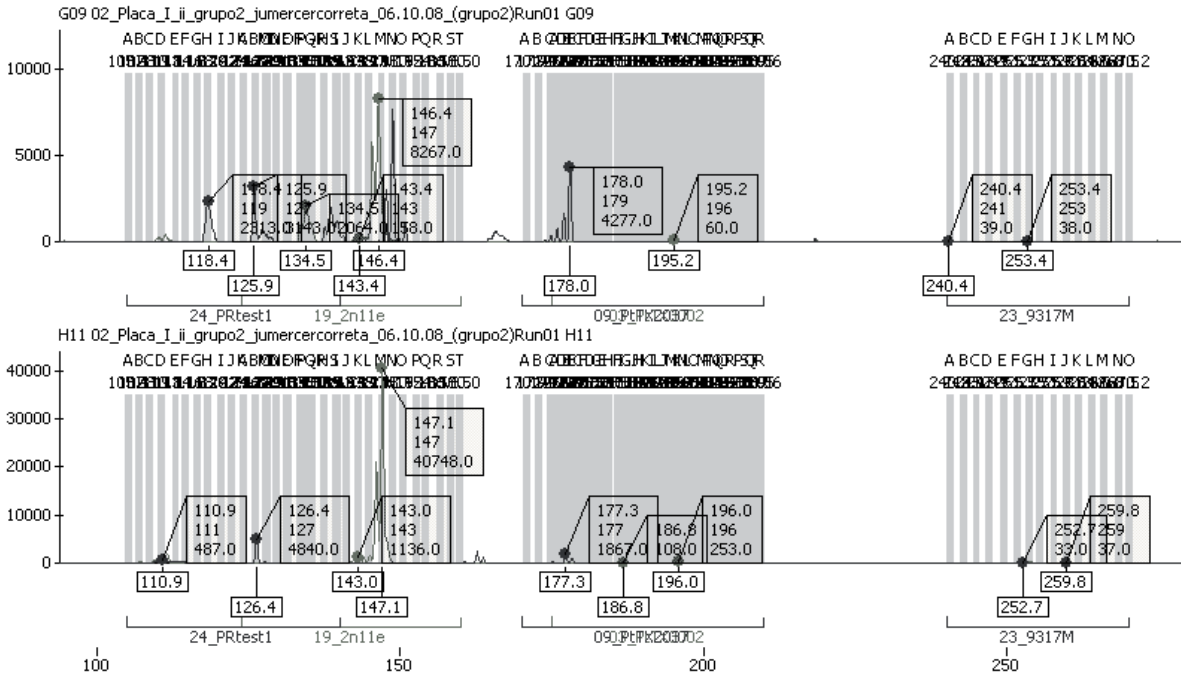


Figura 1. Exemplo de amostras com picos sobrepostos.

Foi observado também que diferentes concentrações na diluição dos primers, interferiram no pico detectado para os *loci* em algumas amostras. Excesso de primers altera a detecção dos fragmentos de modo acentuado. As causas deste excesso podem ocorrer durante os processos de pipetagem. Assim, um procedimento realizado foi a manutenção periódica da calibração das pipetas e o cuidado para exercer uma pressão igual dos dedos durante sua manipulação em todas as amostras. Contudo, erros de excesso ou escassez podem se manifestar. Por esse motivo, foi evitado alterar as pipetas utilizadas durante as atividades, na tentativa de conter essa possível fonte de erro. Durante as atividades de PCR, o próprio experimentador foi responsável pela execução de todo os procedimentos, o que deve ter evitado heterogeneidade dos erros para a obtenção dos produtos de PCR na genotipagem das amostras. Erros de genotipagem devido à pipetagem já foram referenciados em HOFFMAN e AMOS (2005).

Como uma série de fatores e suas interações podem alterar o pico detectado na genotipagem, o software do MegaBace conta com um aplicativo, o Peakfilter, que representa um intervalo de tamanho esperado para a detecção dos fragmentos em *bp* para cada *locus* microssatélite. À medida que a experiência do experimentador aumenta, o intervalo esperado para detectar os fragmentos dos *loci* se torna mais refinado. Assim, a calibração do Peakfilter

se torna cada vez mais eficiente em limpar os fragmentos falsos, até que o procedimento de tipagem alélica se torne cada vez mais automatizado. Contudo, por mais calibrado que estivesse o Peakfilter, ainda foram detectados a presença de falsos alelos na tipagem alélica.

Esses falsos alelos foram verificados pelo seguinte experimento: após ter sido atingida a melhor calibração para o Peakfilter do Fragment Analyser para todos os *loci*, três pessoas foram treinadas por dois dias consecutivos, para realizar uma tipagem alélica da população total, para todos os *loci*, de modo independente. Ou seja, uma pessoa não poderia alterar na avaliação da tipagem alélica da outra. Os resultados foram surpreendentes quando comparados. Para um determinado *locus* por exemplo, um avaliador no final da tipagem alélica havia detectado 23 alelos enquanto outro detectou 8 alelos. Fenômenos semelhantes a este ocorreram para quase todos os *loci*, mostrando que a experiência na avaliação dos alelos, mesmo após uma calibração refinada do Peakfilter, influencia acentuadamente no polimorfismo microssatélite que representa a população.

A dificuldade na tipagem alélica (*allele calling*) também é referenciada na literatura como um problema potencial para marcadores SNPs (GHOSH *et al.*, 1997). Uma conclusão que foi percebida no experimento comentado anteriormente, é que a tipagem alélica é extremamente dependente da qualidade dos dados e da própria experiência do avaliador. Erros de genotipagem devido a esta subjetividade humana no julgamento dos alelos verdadeiros também foram constatados em BONIM *et al.* (2004), que realizaram um procedimento semelhante, usando experimentadores independentes para determinar a tipagem alélica para o marcador AFLP.

Como controle, foram usadas amostras vazias em cada uma das placas de 96 células injetadas no MegaBace. Tal procedimento foi preterido a outras formas de controle, uma vez que as contaminações das amostras durante as manipulações de produtos de PCR e suas diluições estão sujeitas a elevado risco de contaminação. Das 36 placas genotipadas na plataforma, quatro placas apresentaram fragmentos amplificados na amostra vazia. Quando isso ocorreu, as placas não foram invalidadas, mas cautelas mais rigorosas foram aplicadas durante a avaliação da tipagem alélica.

Para estudos de microssatélites, em função da frequência com que podem aparecer alelos nulos (PRIMMER *et al.*, 1995), alguns softwares que permitem detectá-los já estão disponibilizados, sendo um destes o MLTR (RITLAND, 2002). Os alelos nulos podem

ocorrer, gerando a não amplificação de fragmentos para um determinado indivíduo, como também, gerando um excesso de homozigotos para um dado *locus*. Porém, alelos perdidos e excesso de homozigotos, não necessariamente ocorrem por serem alelos nulos, e, é aí que reside a dificuldade de se determinar com confiabilidade, a presença e a indicação dos verdadeiros alelos nulos.

Por essa razão o presente estudo, deu preferência ao método de análise de segregação que foi proposto por GILLET e HATTEMER (1989). Esse método, basicamente consiste em construir zimogramas, ou seja, agrupar cada progênie por *locus*, ordenar os indivíduos da progênie pela ordem de semelhança dos alelos, contar o número de cada grupo de genótipo, e testar a significância da hipótese do modelo de herança do controle genético por teste estatístico testado. Caso a hipótese admitida para o modelo seja não significativo, o genótipo materno é então inferido. A tabela 1 apresenta os possíveis modelos de herança para alelos codominantes e para a presença de alelos nulos. As estimativas de probabilidades para a herança de alelos de cada modelo estão discutidas em detalhes na publicação de GILLET e HATTEMER (1989).

Foi usado o teste G para testar a significância de progênies com excesso de homozigotos. Quando significativo, os alelos suspeitos de serem nulos foram chamados de "0", e "-9" para os alelos perdidos. Dentre os 15 *loci*, alelos nulos que resultaram em significância no teste G foram detectados para dez destes, os quais representaram 66,6% dos *loci*, e 32,7% do total de amostras genotipadas.

Tabela 1. Análise de herança do controle genético de um *locus* para uma única progênie de polinização aberta.

Genótipo Materno	Genótipos Possíveis da Progênie	Hipóteses do Modelo de Herança de Segregação[1]
A_0A_0[2]	A_0A_0 $A_k\_$ $(k \neq 0)$	*Modelo não é apropriado*
A_iA_i $(i \neq 0)$	$A_i\_$ A_iA_k $(k \neq 0, i)$	$N_{ii} \leq N_{ii} + N_{i\_}$
A_iA_0 $(i \neq 0)$	A_0A_0	$N_{00} \leq N_{i\_}$
	$A_i\_$ $A_k\_$ A_iA_k $(k \neq 0, i)$	$N_{k\_} = N_{ik}$ $(k \neq 0, i)$
A_iA_j $(0 \neq i \neq j)$	$A_i\_$ $A_j\_$ A_i A_j	$N_{ij} \leq N_{i\_} + N_{j\_}$
	A_i A_k A_jA_k $(k \neq 0, i, j)$	$N_{ik} = N_{jk}$ $(k \neq 0, i, j)$

[1]/ Proporção esperada do número de genótipos da progênie (N) em relação ao observado.
[2]/ A_0: alelo nulo (ausente); A_i ou A_j: alelo materno possível; A_k alelo paterno.

O teste-G foi o teste de significância utilizado para testar os modelos de herança dos modelos genéticos referenciados na tabela 1, uma vez que o número de indivíduos das progênies era de cinco indivíduos. Este teste estatístico de máxima verossimilhança utiliza a seguinte equação:

$$G = 2\sum_{ij} O_{ij} \cdot \ln(O_{ij}/E_{ij})$$

sendo,

O_{ij}: a frequência observada, E_{ij}: a frequência esperada, e *ln*: o logaritmo natural na base *e* (e = 2,718281828).

Embora a análise de segregação alélica tenha sido efetuada com o propósito de determinar e detectar alelos nulos, tal método se mostrou o mais eficaz para verificar a fidedignidade dos alelos determinados na tipagem alélica, uma vez que além de alelos nulos e perdidos, alelos falsos também foram mais facilmente determinados.

Outra grande vantagem do método de análise de segregação é o fato de que se pode determinar o genótipo de progenitor materno, pois os descendentes necessariamente devem possuir ao menos um dos alelos idênticos ao do progenitor feminino, em progênies de

polinização aberta. Não obstante, a inferência de genótipos não é ambígua quando a progênie é proveniente de cruzamento controlado (LI *et al.*, 2009).

Para verificar se essa ambiguidade estaria prejudicando os genótipos maternos inferidos, 23 genótipos maternos foram efetivamente genotipados para os 15 *loci* e comparados com os seus genótipos inferidos por segregação na progênie. Observou-se uma boa concordância, 93% dos alelos genotipados e inferidos foram coincidentes. Tal resultado aumentou a confiabilidade dos genótipos maternos, inferidos para os 120 progenitores da população selecionada.

Após corrigir manualmente os alelos nulos e falsos e confirmados os alelos perdidos e verdadeiros, um total de quatro tipagens alélicas independentes, já haviam sido realizadas, pelo próprio experimentador. Para detectar a frequência de alelos nulos foi utilizado o software MLTR - versão 3.4, para os dados da quarta tipagem alélica. Os resultados obtidos indicaram que ainda era detectada a presença de alelos nulos em 10 *loci,* numa frequência não superior a 0,02%. Contudo, estes não foram corrigidos, pois sua detecção poderia estar sendo influenciada pela elevada frequência de alelos perdidos.

Para o *locus* 7 foi verificado que ocorria uma duplicação, e esta somente pôde ser confirmada após a análise de segregação, a qual demonstrou que este *locus* apresentava dois grupos de alelos, que segregavam de modo independente nas progênies. Baseado nessa evidência, o *locus* 7 foi dividido em dois *locus* independentes e referenciado como *locus* 7a e 7b. Na publicação de ZHANG e ROSENBERG (2007), são discutidos fenômenos de duplicação em microssatélite. Contudo, não foi encontrado na literatura algum experimento que tenha tratado os *locus* de uma duplicação de modo independente.

Na tentativa de diminuir o número de alelos perdidos no experimento, foram regenotipadas seis placas no MegaBace. Tal procedimento reduziu em média, 16,7% o número de alelos perdidos. Mesmo assim, o total de alelos perdidos ainda permaneceu elevado, com média de 46,27% de amostras perdidas num total de 18,080 dados, que representavam a planilha de dados de 1130 indivíduos genotipados para 16 *loci* (Tab.1.Cap.I).

Para verificar se a genotipagem de alelos coincidia em corridas independentes no MegaBace, foram incluídas 20 amostras escolhidas aleatoriamente na população, que não haviam tido alelos perdidos na primeira genotipagem. Essas amostras foram introduzidas em uma das seis placas, sem realizar nova reação de PCR. Neste experimento, foi detectado que

8% de alelos dos 640 alelos esperados para os 16 *loci* dos 20 indivíduos, não foram coincidentes. Tal resultado poderia estar indicando que pelo menos 8% dos erros estariam associados a atividades de diluição do produto de PCR, ou oscilações da própria plataforma, uma vez que as atividades anteriores não haviam sido refeitas.

Após a regenotipagem, duas novas tipagens alélicas foram efetuadas, totalizando seis tipagens alélicas realizadas para obtenção da planilha final dos dados, que seguiram para as análises bayesianas para caracterização da estrutura genética da população. Após todas estas conferências, aceita-se que os dados sejam os mais fidedignos possíveis para representar o polimorfismo microssatélite da população pesquisada.

A tabela 1 (Cap.I) apresenta os resultados do número de alelos por *locus* que resultaram na primeira e na sexta última tipagem alélica. Nessa tabela, pode ser observado que em média houve uma redução de 37,34% no número de alelos aceitos como verdadeiros, pois o número médio de alelos por *locus* de 8,3 reduziu para 5,2 alelos na sexta e mais experimentada tipagem alélica. A tabela 4 (An.), apresenta os alelos que foram determinados para os *loci* estudados nesta população.

Os resultados da estatística descritiva de cada *locus* após todas as tipagens alélicas, efetuados no software GDA, estão apresentados na tabela 2 (Cap.I). Como os *loci*: 1, 8 e 17 apresentaram um excesso de dados perdidos, estes foram retirados da planilha de dados que seguiu para as análises de estrutura genética.

Concluiu-se que a plataforma automatizada de eletroforese capilar baseada em fluorescência é um método rápido e eficaz para genotipagem de microssatélites em larga escala, contanto que cautelosa análise da tipagem alélica seja efetuada para conter os freqüentes erros detectados em sistema multiplex.

De todos os testes e constatações discutidas até o momento, torna-se pertinente chamar atenção para o fato de que, embora no capítulo I tenha sido feito referência de alguns exemplos de erros de genotipagem microssatélite que podem afetar gravemente estudos genéticos (THOMPSON, 1976; POMPANON *et al.*, 2005; HOFFMAN e AMOS, 2005; BONIN *et al.*, 2004; GAGNEUX *et al.*, 1997a; GAGNEUX *et al.*, 1997b), nenhum destes relacionados com plantas foi encontrado na literatura. Este fato chama atenção, uma vez que erros de genotipagem não devem ser pouco frequentes, e tão pouco, estes erros são restritos

aos marcadores microssatélites. Mesmo assim, suas implicações nos estudos podem ser graves.

Antes de iniciar a discussão sobre os resultados das análises realizadas para caracterizar a estrutura genética da população, um breve comentário se torna pertinente. O termo para determinar a estrutura e o adjetivo de estar estruturada são bastante citados e recebem diferentes conotações em muitas instâncias. Para evitar confusão terminológica, o presente documento considera que analisar a estrutura genética de uma população, não denota necessariamente, que a mesma seja estruturada. Sendo o fato de estar estruturada, sinônimo de estar fragmentada em grupos genéticos distintos.

Para caracterizar a estrutura genética da população selecionada, enfoque foi dado para a inferência bayesiana em análises de agrupamento *multicluster*, após, a validação da fidedignidade dos 15 *loci* microssatélites (SSR) nas 1130 amostras ter sido experimentado. As análises preliminares realizadas com o software STRUCTURE apontaram que a população selecionada é fragmentada em cinco grupos genéticos.

Ao contrário do que se esperava, não houve correspondência entre os cinco grupos genéticos inferidos pelo polimorfismo microssatélite com as cinco florestas plantadas, usadas para sintetizar a população selecionada. Tal resultado surpreendeu, uma vez que as FPs *a priori*, foram assumidas como sendo de cinco origens ou procedências distintas.

Para certificar os resultados obtidos e obter mais informações para a população, foi adotado um procedimento de *learning-samples* ou de conhecimento de amostras, buscando encontrar o número de grupos genéticos (ou K) mais provável. Somente então, os parâmetros genéticos foram inferidos. Neste procedimento de conhecimento de amostras, duas estatísticas auxiliares *ad hoc* foram implementadas para determinar o resultado *a posteriori* mais verossímil, as estatísticas *ad hoc* de PRITCHARD (2000) e de EVANNO (2005).

Após 100 análises terem sido realizadas sem informar a origem das FPs nos dados, usando o modelo *admixture* no STRUCTURE, os resultados da melhor probabilidade *a posteriori*, de acordo com a estatística *ad hoc* de PRITCHARD, continuavam a indicar que cinco grupos genéticos era o valor que melhor representava a população selecionada. Usando esses grupos e, determinando-se como as FPs se distribuíam entre eles, foi possível concluir que estas representavam um mosaico da combinação destes grupos (Fig.1.Cap.II). Foi

observado que o quê as diferenciava era a proporção de contribuição de cada grupo genético na composição de cada FP.

Foram então realizadas análises, informando a origem FP. Foi observado que a diversidade genética entre elas era semelhante, com os resultados obtidos para o número de alelos polimórificos – média de 5,55 alelos por *locus*, e, para a heterozigosidade esperada e a observada. Resultado pouco provável, se considerar que estas populações realmente proviam de origens genéticas distintas. Nestas análises foi retirada a FP_4 por ter um tamanho muito reduzido de indivíduos. Os valores de índice de fixação foram menores para FP_2, sendo esta a mais divergente dentre as FPs (Tab.2.Cap.II).

Para levantar pistas do fluxo gênico entre as FPs foram informadas nas análises a origem FP, mas não foi informado a origem FP para os indivíduos que haviam resultado em *admixutures* (híbridos ou misturas entre grupos genéticos) na análise anterior. Foi então obtido a estimativa para o intervalo de confiança da coancestria destes *admixutures* em relação à origem dos cinco grupos genéticos. Também foram determinadas as probabilidades do número de gerações passadas em que o evento de hibridização deve ter ocorrido, ou seja, na geração parental, de avós ou bisavós. Os resultados indicaram que 60% dos *admixutures* foram gerados na geração parental, o quê coincide com o momento da polinização que ocorreu na geração das FPs para dar origem às sementes que representam a população selecionada.

Para verificar como os *admixutures* interferiam nas análises, todos os indivíduos não *admixutures* foram retirados, bem como, todos os indivíduos da FP_4. A melhor probabilidade *a posteriori* para estes dados, foi de K=3 ao invés do K=5 (Fig.3.Cap.II). As FPs mais distintas foram FP_2 e FP_3, sendo a maioria dos indivíduos da FP_5 um híbrido entre FP_2 e FP_3. Embora metade dos indivíduos provenientes da FP_1 apresentasse híbridos semelhantes aos existentes nas outras FPs, esta população apresentava um grupo genético distinto que ocorria em pequenas proporções nas outras FPs, mas que parecia não intercruzar com os outros grupos genéticos detectados. Entretanto, todas estas constatações não puderam ser testadas de modo mais aprofundado por outras análises, próprias para determinar fluxo gênico, uma vez que não havia informações para a localização geográficas das FPs.

Embora não fosse um objetivo, a análise de segregação por *locus* dentro de progênie permitiu a inferência dos alelos da geração materna da população selecionada, promovendo

respaldo para a interpretação dos resultados das análises de estrutura genética da população de pesquisa. Concluiu-se que a geração materna é mais provável de ter sido plantada pela mistura de três origens distintas de sementes, uma vez que MK=3 foi determinado para esta geração. Comparando-se os coeficientes de endogamia de Wright (F_{ST}), estes foram 2 a 3 vezes maiores nas progênies do que em relação aos da geração materna (Tab.5.Cap.II).

Foi observado que os grupos genéticos detectados na geração materna se mantinham na geração da população selecionada e que estes se hibridizavam de modo desigual. Tal fenômeno pode estar indicando a presença de cruzamentos preferenciais que ocorreram durante a polinização das FPs para formar as sementes da geração da população selecionada. Como a localização geográfica das FPs não foi disponibilizada, testes para confirmar a hipótese de cruzamentos preferenciais dentre estes grupos maternos, não puderam ser realizados. Mesmo assim, há fortes evidências de que estes, provavelmente, ocorreram na população selecionada.

De acordo com a estatística *ad hoc* de EVANNO, o modelo de migração que melhor explicou os resultados foi o de Zona-de-contato ou "*Stepping-stone*". Foi realizada uma pequena modificação na expressão da estimativa *ad hoc* de EVANNO, por isso ao invés do resultado da estatística ser o ΔK original, este passou a ser chamado de ΔK_m. Tal resultado representa o pico de maior valor apresentado graficamente para as probabilidades *a posteriori* de K, estimado com base nos valores das medianas ao invés das médias. O maior valor de ΔK_m foi para K=2 (Fig.5 CapI), o segundo maior para K=4 e o terceiro maior para K=19. Também foram analisados os mesmos dados, alterando o número de *loci* de 15 para 13 nas planilhas de dados, e observou-se que os resultados praticamente não sofreram alteração. As equações utilizadas na modificação realizada para a equação original de ΔK encontram-se no Item 3 do Anexo.

Surpreendente foi o resultado obtido após a verificação de quais eram os indivíduos que faziam parte do K=2 obtido na estatística *ad hoc*. Foi observado que os indivíduos de valor genético *i* ou *ii* se subdividiam nestes dois grupos genéticos com uma coincidência maior do que 90%. Sendo que este resultado foi gerado sem informar o valor genético de nenhum indivíduo. Lembrando que, estes valores *i* ou *ii* designam a melhor (*i*) e a (*ii*) segunda melhor árvore dentro de progênie, em função da seleção para produtividade de madeira que havia ocorrido em teste de progênie previamente na síntese da população selecionada. Já os

outros K=3 dos cinco grupos genéticos, foram parcialmente coincidentes com os mesmos MK=3 determinados para caracterizar a geração materna.

Esses resultados serviram para construir uma hipótese para explicar as causas mais prováveis dos cinco grupos genéticos (K=5) caracterizados na população de progênies. Que estes tenham se formado pela ocorrência de cruzamentos preferenciais entre as três diferentes origens detectadas na geração materna, sendo o fator que mais contribuiu para a forte fragmentação da população foi a pressão de seleção que ocorreu dentro de progênies para selecionar as duas melhores árvores para produtividade de madeira.

Outro resultado que chamou a atenção foi que as árvores de melhor valor genético (*i*) mantiveram uma maior variabilidade genética em comparação ao das árvores de segundo melhor desempenho (*ii*), com valores superiores em heterozigosidade observada e em número de alelos maternos mantidos.

Para verificar como os indivíduos da população se distribuíam entre e dentre os grupos genéticos, após obter o resultados de todas as análises de cada hipótese testada, foram feitas análises de permutações usando o software CLUMMP 1.1.2 (JAKOBSSON e ROSEMBERG, 2007. Estas análises aumentaram a confiabilidade de todos os resultados, uma vez que os indivíduos podem sofrer alterações para a probabilidade de coancestria com os grupos genéticos, inferidos durante as análises bayesianas do STRUCTURE. O CLUMMP permite analisar todas as matrizes de coancestria simultaneamente por permutações e informa a matriz mais verossímil de distribuição dos indivíduos. Para populações complexas, como foi o caso, em que um grande número de indivíduos *admixture*s são detectados, os resultados do CLUMMP são extremamente relevantes.

A figura 7 (Cap.I) apresenta um resultado que permitiu decompor a população total de seleção em 19 subgrupos prováveis, indicados pelo forte pico de ΔK_m em K=19 que resultou da estatística *ad hoc* de EVANNO, usando as permutações do CLUMMP para a indicação dos indivíduos. Essa figura foi determinada usando-se os dados das distâncias genéticas inferidas para cada subgrupo, baseados na divergência de frequências alélicas estimadas no STRUCTURE. Tal imagem facilita a observação visual de como a distribuição das distâncias genéticas entre os 19 subgrupos, se agrupavam em 5 grupos genéticos principais e também, como as árvores *i* e *ii* se distribuíam espacialmente dentre os grupos. Nota-se que o modelo de migração *stepping-stone* se torna evidente com esta imagem e este, aparentemente, sugere que

no mínimo exista uma zona de contato de duas dimensões, a primeira separando as árvores *i* das *ii* e a segunda separando em subgrupos, os grupos destas árvores.

Uma das razões dos esforços da presente pesquisa em priorizar a busca pelo verdadeiro K, ao invés de buscar estimativas para outros parâmetros genéticos, foi o fato de que a fragmentação espúria distorce todos os valores estimados para qualquer parâmetro genético, e, portanto, não são válidos. Dentre os métodos de agrupamento que permitem a busca pelo K mais provável, o mais promissor e o mais discutido na literatura é o disponibilizado pelo software STRUCTURE (ROSEMBERG, *et al.*, 2001; THORNSBERRY *et al.*, 2001; PRITCHARD *et al.*, 2010).

A tabela 4 (Cap.II) apresenta os resultados de F_{ST} e He para as árvores *i* e *ii* em função dos cinco grupos genéticos determinados. Note que, em média, os valores de F_{ST} para as árvores *i* foram de 0,24 e He 0,45 e para as árvores *ii*, os valores de F_{ST} e He foram de 0,36 e 0,4, respectivamente.

Alguns estudos relatam que a heterozigose de um indivíduo, determinada por marcadores neutros tais como os microssatélites, pode estar correlacionada com alguns componentes de adaptabilidade, tais como: a fecundidade (AMOS *et al.*, 2001), a resistência a doenças (COLTMAN *et al.*, 1999) e o sucesso na vida reprodutiva (SLATE *et al.*, 2000). O fato de a maior heterozigosidade ter sido detectada para árvores *i* é intrigante, visto que o volume de madeira não parece ser uma característica adaptativa. Por outro lado, o grupo com as árvores *ii* apresentou o maior coeficiente de endogamia (F_{ST}) em média.

O maior coeficiente F_{ST} poderia invocar a idéia de que o menor desempenho em volume de madeira em comparação com árvores *i*, poderia estar atribuído à depressão endogâmica como o mecanismo responsável. Quando um genoma carrega alelos recessivos, os indivíduos com homozigose média elevada apresentarão uma tendência de serem menos adaptados e de apresentar um menor desempenho para a maioria das características, sendo esta, a base da depressão por endogamia clássica (KELLER e WALLER, 2002).

Outra importância que se deve perceber da metodologia aplicada de conhecimento de amostras, é que a mesma permitiu conhecer e investigar informações que não teriam existido, caso as hipóteses não tivessem sido levantadas na medida em que os resultados dos testes apontavam determinado caminho de investigação. Os resultados preliminares da não

correspondência entre as FPs para com os grupos genéticos, conforme era acreditado, foi o fator responsável em promover a busca pela metodologia proposta.

Os pomares clonais estabelecidos por programas de seleção, normalmente na primeira geração de seleção, são sintetizados por seleção fenotípica (massal) para explorar a variabilidade genética entre e dentre populações plantadas em reflorestamentos comerciais. Essa informação, não se encontra em publicações, mas é referenciada no meio do Melhoramento Florestal para esta espécie. Outra informação que também é discutida neste meio é de que a qualidade de sementes provavelmente não deve ter sido enfatizada há cerca de 30 anos atrás, no momento em que as FPs do presente estudo foram estabelecidas. Isso leva a acreditar que estas FPs não necessariamente apresentariam uma fonte única de origem geográfica de sementes.

GAO *et al.* (2007) estendeu a abordagem bayesiana do popular STRUCTURE (PRITCHARD *et al.*, 2000) para a inferência simultânea das taxas de consanguinidade ou de autofecundação e para a classificação da população por origem genética usando análises *multilocus*. Isto é conseguido quando se deixa de assumir o equilíbrio de Hardy-Weinberg nos grupos e, assim, estima-se as freqüências genotípicas esperadas com base nas taxas de endogamia e endocruzamento. Os pesquisadores demonstraram a necessidade de tal procedimento quando verificaram que a autofecundação induz sinais espúrios de fragmentação para a população, quando esta é analisada pelo critério de equilíbrio de HW, com uma tendência também para sinais espúrios de *admixture*s.

A fragmentação de uma população pode gerar desequilíbrio de ligação entre distantes regiões cromossômicas e, em estudos de associação, isso pode levar a associações espúrias entre as características e os polimorfismos, a menos que a estrutura da população esteja bem caracterizada por análises estatísticas. Populações de plantas em geral apresentam estruturas genéticas geralmente tidas como de difícil resolução imediata, principalmente por ser comum a ocorrência de acasalamento entre parentes e a hibridização de diferentes origens (CAMUS-KULANDAIVELU *et al.*, 2007). GAO *et al.* (2007) mediram o desempenho do método proposto pelo INSTRUCT, com extensas simulações coalescentes e demonstraram que a abordagem pode corrigir esse viés.

O STRUCTURE gera agrupamentos baseado em ambos o desequilíbrio de Hardy-Weinberg (HW) e no desequilíbrio de ligação (LD) transiente, causado pela presença de

admixutures entre populações. Basicamente, a análise busca agrupar os indivíduos em grupos, até que ambos o desequilíbrio de HW e de LD se minimizem. Portanto, a presença de LD nos dados melhora os resultados de agrupamento (FALUSH *et al.*, 2003). Por outro lado, um LD excessivo ou um desvio forte ao equilíbrio de HW pode levar a uma superestimação do número de grupos detectados (FALUSH *et al.*, 2003). Como quase todos os *locus* microssatélites apresentaram desvios ao equilíbrio de HW e alguns LD, foi esperado que os cinco grupos genéticos pudessem estar sendo superestimados. Se os cinco K estivessem superestimados, seria esperado que o software INSTRUCT detectasse um menor número de grupos genéticos. Contudo, esta expectativa não ocorreu e, na realidade, o resultado para o número mais provável de K foi idêntico para ambos os critérios testados.

O coeficiente de endogamia variou de 0,19 a 0,42 nas análises do STRUCTURE para os cinco grupos genéticos, em 1.130 amostras e 15 *loci*. Na análise assumindo endogamia do INSTRUCT, os valores de F_{ST} variaram de 0,022 a 0,227 para os mesmos K=5, mas nestas análises apenas 738 indivíduos e 10 *loci* puderam ser analisados, uma vez que o INSTUCT não permitia tantos dados perdidos como o STRUCTURE – foram excluídos os *loci*: 01, 06, 08, 17 e 21. A tabela 7 (Cap.II) resume estes resultados. Concluiu-se que os valores de endogamia foram menores assumindo-se a endogamia, do que assumindo-se o equilíbrio de Hardy-Weinberg. Contudo, houve consistência nos grupos genéticos detectados e na distribuição dos indivíduos entre e dentre esses grupos, sob ambos os critérios assumidos nas análises bayesianas.

Como a estimação dos parâmetros genéticos variou bastante conforme o método, e como foi observado que nas análises realizadas para cada grupo genético em separado, um número menor do que 180 indivíduos gerava instabilidade para os resultados obtidos, sugere-se que para uma definição mais verossímil na inferência dos parâmetros genéticos, faz se necessário o aumento do número de *loci* microssatélites para essa população.

A figura 8 (Cap.II) apresenta o resultado dos cinco grupos genéticos mais prováveis detectados pelo INSTRUCT, graficamente expostos em ordem crescente de endogamia, em função dos K=5 previamente detectados no STRUCTURE. A Tabela 6 (Cap.II) resume os resultados obtidos, para o número de indivíduos que contribuíram para a formação de cada grupo genético no INSTRUCT em função dos grupos genéticos do STRUCTURE. Foi determinado que o grupo com o menor coeficiente de endogamia, o cluster 1 do INSTRUCT com F_{ST} de 0,022 (Fig.8.Cap.II), foi formado com a contribuição de 56% dos indivíduos

pertencentes ao K_4 resultante do STRUCTURE. O K_4 é quase exclusivamente formado por árvores *i* (Fig.7.Cap.II). Quanto ao grupo de maior coeficiente de endogamia, o cluster 5 do INSTRUCT com F_{ST} de 0,227, foi formado pela contribuição de 72% dos indivíduos do K_2 do STRUCTURE. O K_2 é quase exclusivamente formado por árvores *ii*. Mais uma vez é pertinente enfatizar que para as análises realizadas no INSTRUCT, também não tinham sido informadas a classificação *i* ou *ii* dos indivíduos.

Em CAVALLI-SFORZA *et al.* (1994), bem como em YU *et al.* (2009), os valores de F_{ST} de 0,01 a 0,05 seriam os valores esperados para populações de *Pinaceaes* e de espécies de polinização aberta em condições naturais; já os valores de 0,05 a 0, 30 seriam os valores esperados para populações divergentes e fragmentadas. Estes valores, comparados com os resultados obtidos, indicam que a população selecionada apresenta alto grau de fragmentação.

Embora os valores estimados para os parâmetros genéticos tenham variado, de qualquer forma a fragmentação detectada implica num risco de endogamia que é incompatível para uma primeira geração de seleção, caso o objetivo do programa florestal vise a exploração da população para o seleção intrapopulacional em longo prazo.

CONCLUSÕES

1. A utilização dos *loci* microssatélites utilizados e os métodos aplicados para sua detecção e validação permitem caracterizar a estrutura genética da população selecionada com uma razoável confiabilidade dos resultados obtidos.

    1.1 A validação da fidedignidade do polimorfismo microssatélite permite concluir que o procedimento de genotipagem para um tamanho de amostras de 1,130 indivíduos é necessária, uma vez que foi detectada a presença de alelos perdidos, nulos e falsos misturados aos alelos verdadeiros. Uma vez não considerada, poderia comprometer a confiabilidade dos resultados obtidos.

    1.2 A análise de segregação alélica por *locus* dentro de progênies utilizando-se microssatélites é uma metodologia que permite validar a fidedignidade do polimorfismo alélico representativo da população selecionada.

    1.3 A análise de segregação por *locus* dentro de progênie permite inferir os alelos da geração materna da população selecionada e promove respaldo para a interpretação dos resultados das análises da estrutura genética da população selecionada.

2 Comparando os dois métodos de análises bayesianas, assumindo equilíbrio de Hardy-Weinberg e assumindo endogamia, percebe-se que há consistência para os números de grupos genéticos detectados e para a distribuição dos indivíduos entre e dentre esses grupos, sob ambos os modelos de análises. Contudo, é verificado que os valores dos parâmetros genéticos definidos nos dois modelos para coeficiente de endogamia de Wright (F_{ST}) e para heterozigosidade esperada (He) são acentuadamente distintos.

3 As análises bayesianas apontam que a população selecionada é fragmentada em cinco grupos genéticos. Não havendo correspondência entre os cinco grupos genéticos inferidos pelo polimorfismo microssatélite com as cinco florestas plantadas usadas para sintetizar a população selecionada.

    3.1 A proposta metodológica de análises de conhecimento de amostras permite verificar a consistência da fragmentação detectada e permite propor uma

explicação biológica para a existência dos cinco grupos genéticos definidos na população selecionada.

4 A explicação biológica mais coerente para os resultados obtidos consiste, em um primeiro momento, na formação de dois fragmentos populacionais mais divergentes. Estes foram decorrentes do efeito da seleção dentro de progênies para volume de madeira. Já, os outros três fragmentos existentes, são prováveis manifestações de três origens maternas geneticamente distintas, que se mantiveram na geração da população selecionada. Ou seja, podem indicar a existência de bases genéticas distantes que não se homogeneizaram para formar a primeira geração da população recombinante.

4.1 As árvores de melhor valor genético para volume de madeira (*i*) dentro das progênies da população selecionada apresentam menor coeficiente de endogamia e maior heterozigosidade do que as de segundo valor genético (*ii*), sendo o fenômeno da endogamia clássica a melhor explicação para esse resultado.

RECOMENDAÇÕES

1. Sugere-se que, para uma definição mais verossímil na inferência dos parâmetros genéticos, se faz necessário o aumento do número de *loci* microssatélites aplicados nessa população.

2. A fragmentação detectada implica em um maior risco de endogamia para o avanço das gerações de seleção em longo prazo.

3. A definição da estrutura genética da população alvo de melhoramento intrapopulacional se mostra uma boa ferramenta para monitorar a variabilidade genética nas próximas etapas de avanço das gerações de seleção.

REFERÊNCIAS BIBLIOGRÁFICAS

ALLONA, I.; QUINN, M.; SEDEROFF, R.; WHETTEN, R.W.; *et al.* Analysis of xylem formation in pine by cDNA sequencing. **Proc. Natl. Acad. Sci**., n. 95, p. 9693-9698. 1998.

AL-RABAB'AH, M.A.; WILLIAMS, C.G. Population dynamics of *Pinus taeda* L. based on nuclear microsatellites. **Forst. Ecol. Manag**., n. 163, p. 263-271. 2002.

AMOS, W.; WILMER, J.; FULLARD, K. *et al.* The influence of parental relatedness on reproductive success. **Proc. Roy. Soc. Lond**., n. 268, p. 2021–2027. 2001.

BANKS, M.A.; EICHERT, W. WHICHRUN. Version 3.2: a computer program for population assignment of individuals based on *multilocus* genotype data. **J. Heredity**, n. 91, p.87-89. 2000.

BONIN, A.; *et al*: How to track and assess genotyping errors in population genetics studies. **Mol. Ecol.**, n. 13, p. 3261–3273. 2004.

BROWN, G.R.; BASSONI, D.L.; GILL, G.P.; FONTANA, J.R.; *et al.* Identification of quantitative trait *loci* influencing wood property traits in loblolly pine (Pinus taeda L). III. QTL verification and candidate gene mapping. **Genetics,** v.164, p.1537-1546, 2003.

BROWN, G.R.; GILL, G.P.; KUNTZ, R.J. LANGLEY, C.H. NEALE, D.B. Nucleotide diversity and linkage disequilibrium in loblolly pine. **Proc. Nat. Acad. Science,** n.101, p.15255-15260. 2004.

BUCKLER, IV.E.S.; THORNSBERRY, J.M. Plant molecular diversity and applications to genomics. **Curr. Opin. Plant Biol,** v.5, p.107-111, 2002.

BURDON, R. D. & SHELBOURNE, C. J. A. - Breeding populations for recurrent selection: conflicts and possible solutions. In: BURLEY, J. & NIKLES, D. G. Selection and breeding to improve some tropical conifers. **Comm. Forestry Inst.**, v.2, p. 408-429. 1973.

CAMUS-KULANDAIVELU, L., VEYRIERAS, J.B.; GOUESNARD, B.. CHARCOSSET, A.; MANICACCI, D. Evaluating the reliability of structure outputs in case of relatedness between individuals. **Crop Science**, n. 47, p. 887–892. 2007.

CAVALLI-SFORZA, LL.; MENOZZI, P.; PIAZZA, A. **The History and Geography of Human Genes.** Princeton University Press, Princeton, NJ. 1994. 163p.

CHAGNÉ, D.; LALANE, C.; MADUR, D.; KUMAR, S.; PLOMION, C.A. *et al.* High density genetic map of maritime pine based on AFLPs. **Ann For Sci,** v,.59, p.627-636. 2002.

CHAGNÉ, D.; BROWN, G.; LALANNE, C., MADUR, D.; POT, D.; NEALE, D.; PLOMION, C. Comparative genome and QTL mapping between maritime and loblolly pines. **Mol Breed**, v.12, p.185–195. 2003.

CHAGNÉ, D.; ECHT, C.; RICHARDSON, T.; PLOMION, C.; *et al.*: Cross-species transferability and mapping of genomic and cDNA SSRs in pines. **J. Theor. Appl. Genet**., n. 109, p. 1204-1214. 2004.

COCKERHAM, C.C. Variance of gene frequencies. **Evolution**, n. 23, p. 72-84. 1969.

COLTMAN, D.; PILKINGTON, J.; SMITH, J.; PEMBERTON, J. Parasite mediated selection against inbred Soay sheep in a free-living, island population. **Evolution**, n. 53, p. 1259–1267. 1999.

CORANDER, J.; WALDMANNB, P.; SILLANPÄÄ, M.J. Bayesian Analysis of Genetic Differentiation Between Populations. **Genetics**, n. 163, p. 367-374. 2003.

CROW, J.F.. Hardy, Weinberg and language impediments. Genetics, n. 152, p. 821-825. 1999.

DOYLE, J.J.; DOYLE, J.L. Isolation of plant DNA from fresh tissue. Focus, n. 12, p. 13-15. 1990.

ELSIK, C.G.; MINIHAN, V.T.; Hall, S.E.; SCARPA, A.M.; WILLIAMS, C.G. Low-copy microsatellite markers for *Pinus taeda* L. **Genome**, n. 43, p. 550–555. 2000.

EVANNO, G.; REGNAUT, S.; GOUDET, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. **Molecular Ecology**, n. 10, p. 1365-1375. 2005.

EWEN, K.R. *et al.* Identification and analysis of error types in high-throughput genotyping. **Am. J. Hum. Genet**., n. 67, p. 727–736. 2000.

FALUSH, D.; STEPHENS, M.; PRITCHARD, J.K. Inference of population structure using *multilocus* genotype data: linked *loci* and correlated allele frequencies. **Genetics**, n. 164, p. 1567–1587. 2003.

FALUSH, D.; STEPHENS, M.; PRITCHARD, J.K. Inference of population structure using *multilocus* genotype data: dominant markers and null alleles. **Molecular Ecology**, n.35, p.895-908. 2007.

FAO. Sociedade Brasileira De Silvicultura. **Fatos e Números do Brasil Florestal**. São Paulo, s.d. Dezembro, 110p. 2007.

FISHER, R. A., 1930 **The Genetical Theory of Natural Selection**. Clarendon Press, Oxford. 318p.

FRANKHAM, R.; BALLOU, J.D.; BRISCOE, D.A. **Fundamentos de Genética da Conservação**. Ribeirão Preto, SP: SBG (Sociedade Brasileira de Genética), 2008. 280 p.: il. – Título em inglês: A primer of Conservation Genetics. – Traduzido para o português por Mercival Roberto Francisco e Izeni Pires /Farias. ISBN: 978-85-89265-08-9.

GAGGIOTTI, O.E. *et al.* Patterns of colonization in a metapopulation of grey seals. **Nature**: n. 13, p. 424-427. 2002.

GAGNEUX, P.; BOESCH, C.; WOODRUFF, D.S. Microsatellite scoring errors associated with noninvasive genotyping based on nuclear DNA amplified from shed hair. **Mol.Ecol.**, n.6, p. 861–-868. 1997a.

GAGNEUX, P.; WOODRUFF, D.S.; BOESCH, C. Furtive mating in female chimpanzees. **Nature**, n.387, p.:358–359. 1997b.

GAO, H.; WILLIAMSON, S.; BUSTAMANTE,C.D. A Markov Chain Monte Carlo Approach for Joint Inference of Population Structure and Inbreeding Rates From *Multilocus* Genotype Data. **Genetics**, n.176, p.1635-1651. 2007.

GELFAND, A.E.; SMITH, A.F.M. Sampling-based approaches to calculating marginal densities. **J. Amer. Statist. Assoc**., n. 85, p. 398-409. 1990.

GHOSH, S.; *et al.* Methods for precise sizing, automated binning of alleles, and reduction of error rates in large-scale genotyping using fluorescently labeled dinucleotide markers. **Genome Res.**, n. 7, p. 165–178. 1997.

GILLET, E.; HATTEMER, H.H. Genetic analysis of isoenzyme phenotypes using single tree progenies. **Heredity**, n.63, p. 135-141. 1989.

GOLFARI, L. **Coníferas aptas para reflorestamento nos Estados do Paraná, Santa Catarina e Rio Grande do Sul**. Brasil Florestal: Boletim Técnico, Brasília, n. 1, p. 1-71, out., 1971.

GONZÁLEZ-MARTÍNEZ, S.C.; ERSOZ, E.; BROWN,G.R.; WHEELER, N.C.; NEALE, D.B. DNA Sequence Variation and Selection of Tag Single-Nucleotide Polymorphisms at Candidate Genes for Drought-Stress Response in Pinus taeda L. **Genetics,** v.172, p.1915-1926, Mar. 2006 a.

GONZÁLEZ-MARTÍNEZ, S.C.; KRUTOVSKY, K.V; NEALE, D.B. Forest-tree population genomics and adaptive evolution. **New Phytologist**, n. 170, p. 227-238. 2006.

GONZÁLEZ-MARTÍNEZ, S.C.; WHEELER, N.C.; NEALE, D.B. *et al.* Association Genetics in Pinus taeda L. I. Wood Property Traits. **Genetics,** Published Articles Ahead of Print, published on November 16, 2006 as 10.1534/Genetics.106.061127. 2006 b.

GROTKOPP E, REJMANEK M, SANDERSON MJ, ROST TL. Evolution of genome size in pines (Pinus) and its life-history correlates: Supertree analyses. **Evolution**, n. 58, p. 1705-1729. 2004.

HALDER I, SHRIVER M. Measuring and using admixture to study the genetics of complex diseases. **Hum Genet.**: n. 1, p.52-62. 2003.

HALLAUER AR, MIRANDA JBF. **Quantitative genetics in maize breeding,** 2nd edn. Ames: Iowa State University Press; 1988.

HIRSCHHORN, J. N., DALY, M.J. Genome-wide association studies for common diseases and complex traits. **Nature Rev. Genet**., v.6, p.95-108. 2005.

HOFFMAN, J.I.; AMOS, W. Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. **Mol. Ecol.**, n. 14, p. 599–612. 2005.

HOLSINGER, K.E.; WEIR, B.S. Genetics in geographically structured populations: defining, estimating and interpreting FST**. Nature Rev. Gen.**, n.10, p. 639-650. Sept., 2009.

HUBISZ, M.J.; FALUSH, D.; STEPHENS, M.; PRITCHARD, J.K. Inferring weak population structure with the assistance of sample group information. Mol. Ecol. Res., n. 9, p. 1322-1332, 2009.

JAKOBSSON, M.; ROSENBERG, N.A. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. **Bioinformatics**, n. 14, p. 1801–1806. 2007.

KELLER, L.; WALLER, D. Inbreeding effects in wild populations. **T. Ecol. Evol**., n. 17, p. 230–241. 2002.

KUTIL, B.L.; WILLIAMS, C.G. Triplet-repeat microsatellites shared among hard and soft pines. **J. Hered.**, n. 92, p. 327-332. 2001.

LI, L.; GUO, X.; ZHANG, G. Inheritance of 15 microsatellites in the Pacific oyster *Crassostrea gigas*: segregation and null allele identification for linkage analysis. **Chin. Journ. Ocean. Limnol.**, n. 27, p. 74-79. 2009.

LÜBBERSTEDT, T.; MELCHINGER, A.E.; DUßLE, C.; VUYLSTEKE, M.; KUIPER, M. Relationships among early European maize inbreds: IV Genetic diversity revealed with AFLP markers and comparison with RFLP, RAPD, and pedigree Data. **Crop Sci**., n. 40, p.783-791. 2000.

MAENHOUT, S.; De BAETS, B.; HAESAERT, G. Marker-based estimation of the coefficient of coancestry in hybrid breeding programmes. **Theor Appl Genet**., n.: 118, v.6, p.1181-1192. Apr, 2009.

MAGNUSON, V.L.; *et al.* Substrate nucleotide-determined non-templates addition of adenine by Taq DNA polymerase: implications for PCR-based genotyping and cloning. **BioTechniques**, n. 21, p. 700–709. 1996.

MALÉCOT, G. **Les mathématiques de l'hérédité**. Masson & Cie, Paris. 1948.

POMPANON, F.; BONIN, A.; BELLEMAIN, E.; TABERLET, P. Genotyping errors: causes, consequences and solutions. **Nature Reviews**, n. 6, p.847-859. 2005.

POT, D.; CHANTRE, G.; ROZENBERG, P.; PLOMION, C.; *et al.* Genetic control of pulp and timber properties in maritime pine (Pinus pinaster Ait). **Ann For Sci**, v.59, p.563-575, 2002.

POT, D.; McMILLAN, L.; LePROVOST, G.; PLOMION, C.; *et al.* Nucleotide variation in genes involved in wood formation in two pine species. **New Phytol**., v.167, p.101-112, 2005.

POT, D.;PLOMION, C.; *et al.* QTLs and candidate genes for wood properties in maritime pine (*Pinus* pinaster Ait.). **Tree Genetics & Genomes**, v. 2, p.10-24. 2006.

PRIMMER, C.R.; MOLLER, A.P. ELLEGREN, H. Resolving genetic relationships with microsatellite markers: a parentage testing system for the swallow *Hirundo rustic*. **Mol. Ecol.**, n.4, p. 493-498. 1995.

PRITCHARD, J.K; STEPHENS, M.; DONNELLY, P.J. Inference of population structure using *multilocus* genotype data. **Genetics**, n. 155, p. 945-959. 2000.

PRITCHARD, J. K.; W. WEN, **Documentation for Structure software version 2.** Department of Human Genetics. University of Chicago, Chicago, IL (available at <http://pritch.bsd.uchicago.edu>), 2004.

PRITCHARD, J.K.; WEN, X.; FALUSH, D. **Documentation for structure software: Version 2.3**. Software from http://pritch.bsd.uchicago.edu/structure.html. Feb., 2010.

RITLAND, K. Extensions of models for the estimation of mating systems using *n* independent *loci*. **Heredity**, n. 88, p. 221-228. 2002.

ROSENBERG, N.A. *et al.* Empirical evaluation of genetic clustering methods using *multilocus* genotypes from 20 chicken breeds. **Genetics**, n. 132, p. 699-713. 2001.

ROSENBERG, N. A., PRITCHARD, J. K., *et al.* Genetic structure of human populations. **Science**, v.298, p. 2381-2385. 2002.

ROZENBERG, P.; CAHALAN, C. Spruce and wood quality: Genetic Aspects (A Review). **Silvae Genetica,** v.46, n.5, p.270-274. 1997.

SAMANTA, S., LI, Y.J. ; WEIR, B.S. Drawing inferences about the coancestry coefficient. **Theor. Pop. Biology**, n. 75, p. 312-319. 2009.

SBS - Sociedade Brasileira De Silvicultura. **Fatos e Números do Brasil Florestal**. São Paulo, s.d. Dezembro, 2008. 93p.

SEWELL, M. M., SHERMAN, B. K.. NEALE, D.B. A consensus map for loblolly pine (Pinus taeda L.). I. Construction and integration of individual linkage maps from two outbred three-generation pedigree. **Genetics,** v.151, p. 321-330, 1999.

SEWELL, M.M, BASSONI, D.L; *et al.* Identification of QTL influencing wood properties traits in loblolly pine (Pinus taeada L.). I. Physical wood properties. **Theor Appl Genet**, v. 101, p.1273-1281, 2000.

SEWELL, M.M.; DAVIS, M.F. *et al.* Identification of QTLs influencing wood property traits in loblolly pine (Pinus taeda L.). II. Chemical wood properties. **Theor Appl Genet.**, v. 104, (2-3), p. 214–222, 2002.

SHELBOURNE, C. J. A. **Planning breeding programs for tropical conifers grown as exotics**. In: BURLEY, J & NIKLES, D. G. - Selection and breeding to improve some tropical conifers. Oxford. Commonwealth Forestry Institute, 1973. v. 2 p. 155-179.

SHIMIZU, J.Y.. **Pinus na Silvivultura Brasileira**. Artigo Embrapa Florestas. Página Eletrônica: Ambiente Brasil. Acessado em 31/01/2011. Disponível em: <http://ambientes.ambientebrasil.com.br/florestal/artigos/pinus_na_silvicultura_brasileira.html>.

SLATE, J.; KRUUK, L.; MARSHALL, T.; PEMBERTON, J.; CLUTTON-BROCK, T. Inbreeding depression influences lifetime breeding success in a wild population of red deer (cervus elaphus). **Proc. Roy. Soc. Lond**., n. 267, p. 1657–1662. 2000.

TABERLET, P.; *et al.* Reliable genotyping of samples with very low DNA quantities using PCR. **Nucleic. Acids. Res.**, n. 24, p. 3189–3194. 1996.

THOMPSON, E.A. A paradox of genealogical inference. **Adv. Appl. Probab.**, n. 8, p.648–-650. 1976.

THORNSBERRY, J.M.; GOODMAN, M.M.; DOEBLEY, J. *et al.* Dwarf8 polymorphisms associate with variation in flowering time. **Nature Genetics**, n. 28, p.286–289. 2001.

VAN INGHELANDT, D.; MELCHINGER, A.E.; LEBRETON, C.; STICH, B. Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. **Theor Appl Genet**.: n.120, v.7, p.1289-1299. 2010.

VIGNAL, A.; MILAN, D.; SANCRISTOBAL, M.; EGGEN, A. A review on SNP and other types of molecular markers and their use in animal genetics. **Genet Sel Evol**.: n. 34, p. 275-305. 2002.

WAKAMIYA, I.; NEWTON, R.J.; JOHNSTON, J.S.; PRICE, H.J. Genome size and environmental factors in the genus Pinus. **Am J Bot**, v.80, p.1235-1241. 1993.

WANG, W. Y. S., BARRATT, B. J.; *et al.* Genome-wide association studies: theoretical and practical concerns. **Nat. Rev. Genet**., v.6, p.109-118. 2005.

WAPLES, R.; GAGGIOTTI, O. INVITED REVIEW: What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. **Molecular Ecology**, v. 15, I.:6, p. 1419-1439. 2006.

WEIR, B.S.; COCKERHAM, C.C. Mixed self- and random-mating at two *loci*. **Genetical Research**: n. 21, p. 247-262. 1973.

WEIR, B.S.; COCKERHAM, C.C. Estimating *F*-statistics for the analysis of population structure. **Evolution**, n.38, p. 1358-1370. 1984.

WEIR, B.S.; HILL, W.G. Estimating F-statistics. **Ann. Rev. Genetics**, n. 36, p. 721-750. 2002.

WRIGHT, S. Evolution in Mendelian populations. **Genetics**, n.16, p. 97-159. 1931.

WRIGHT, S. The genetical structure of populations. **Ann. Eugenics**, n. 15, p. 323-354. 1951.

WRIGHT, S. The interpretation of population structure by F-statistics with special regard to system mating. **Evolution**, n. 19, p. 395–420. 1965.

YAZDANI, R., I. SCOTTI, *et al.* Inheritance and diversity of simple sequence repeat (SSR) microsatellite markers in various families of Picea abies. **Hereditas**, v.138, p. 219-227, 2003.

YU, J., S. N. HU, J. WANG, G. K. S. WONG, S. G. LI, *et al.*, A draft sequence of the rice genome (Oryza sativa L. ssp indica). **Science**, v. 296, p. 79-92, 2002.

YU, J., PRESSOIR, G.; BRIGGS, W.H.; VROH Bi, I. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. **Nature Genetics**, n. 38, p. 203-208, 2006.

YU, J.; ZHANG, Z.; ZHU, C. Simulation Appraisal of the Adequacy of Number of Background Markers for Relationship Estimation in Association Mapping. **The Plant Genome**, n.1, v.2, p.63-77. 2009.

ZENG, Z.B.; KAO, C.H.; BASTEN, C.J. Estimating the genetic architecture of quantitative traits. **Genet Res Camb**., v.74, p.279-289. 1999.

ZHANG, K.; ROSENBERG, N.A. On the genealogy of a duplicated microsatellite. **Genetics**, n. 177, p. 2109–2122. 2007.

ZHOU, Y.; GWAZE, D.P.; REYES-VALDÉS, M.H.; BUI, T.; WILLIAMS, C.G. No clustering for linkage map based on low-copy and undermethylated microsatellites. **Genome**, n. 46, p. 809–816. 2003.

ZHU, C.; GORE, M.; BUCKLER, E.S.; YU, J. Status and Prospects of Association Mapping in Plants. **The Plant Genome**: v.1, n.1, p.5-20. 2008.

ZHU, M.; YU, M.; ZHAO, M. Understanding Quantitative Genetics in the Systems Biology Era. Int. **Jour. Biol. Scienc**., n.5, v.2, p.161-170. 2009.

ZOBEL, B. J.; WEIR, R. J. & JETT, J. B. **Breeding methods to produce progeny for advanced - generation selection and to evaluate parent trees**. In: BURLEY, J. & NIKLFS, D.G. Selection and breeding to improve some tropical conifers. Oxford, Commonwealth Forestry Institute, 1973. v. 2 p. 180-202.

ZOBEL, B.; VAN BUIJTENEN, J.P. **Wood variation. Springer Series in Wood Science**. Springer, Berlin Heidelberg New York, p 363, 1989.

ZOBEL, B. J., JETT, J.B. **Genetics of Wood Production**. Springer-Verlag, Berlin/Heidelberg/New York, 1995.

ANEXO I

1. POPULAÇÕES DE ESTUDO

As populações de melhoramento descritas a seguir que foram utilizadas na presente pesquisa fazem parte de um amplo programa do Setor Privado da região de Santa Catarina – Brasil. O projeto é financiado pela iniciativa privada e pelo agente financiador FINEP, sendo o LAMEF (Laboratório de Genética e Melhoramento Florestal do Departamento de Ciências Florestais – UFPR) o órgão executor principal do projeto, cujo responsável e coordenador é o Prof. Dr. Antonio Rioyei Higa – LAMEF/DECIF/UFPR.

Florestas Plantadas [FPs]: representa a população base com 100.000 árvores, onde as sementes de 120 matrizes, que representam a geração materna (M) de cada progênie de polinização aberta, foi fenotipicamente selecionada para parâmetros de produtividade de madeira (Tabela 1). A seleção ocorreu em cinco Florestas Plantadas (FPs), no sul do Brasil – posições geográficas não informadas. Amostras de 23 Ms foram disponibilizadas para o presente estudo.

População Base [PB]: representa a população base de 1ª. geração onde as sementes das FPs foram plantadas em delineamento de blocos casualizados, contendo 120 progênies de polinização aberta instaladas em 5 repetições com 5 plantas por parcela, totalizando 3.000 árvores. Esta população representa o teste de progênie do Programa de Melhoramento.

População Selecionada ou *Breeding Population* [BP]: representa a primeira etapa de seleção da população base, com base no valor genético dos indivíduos aos 12 anos de idade (2009), resultante do teste de progênie para características simultâneas medidas em níveis independentes para volume de madeira. A BP é formada pela seleção de duas árvores dentro de parcelas das 120 progênies de polinização aberta, totalizando um tamanho de 1.130 indivíduos.

População Selecionada de segunda geração: representa a segunda etapa de seleção da população base de 2ª. Geração. Esta seleção leva em conta a diversidade e a divergência genética dos indivíduos determinados no estudo de caracterização genética da estrutura

genética da população BP, e no valor fenotípico para características que envolvem qualidade de madeira. Das 120 progênies de polinização aberta, 80 famílias de maior valor genético para volume de madeira terão um indivíduo selecionado, com base nos critérios mencionados, de modo que a população final irá conter de 80 a 160 indivíduos selecionados.

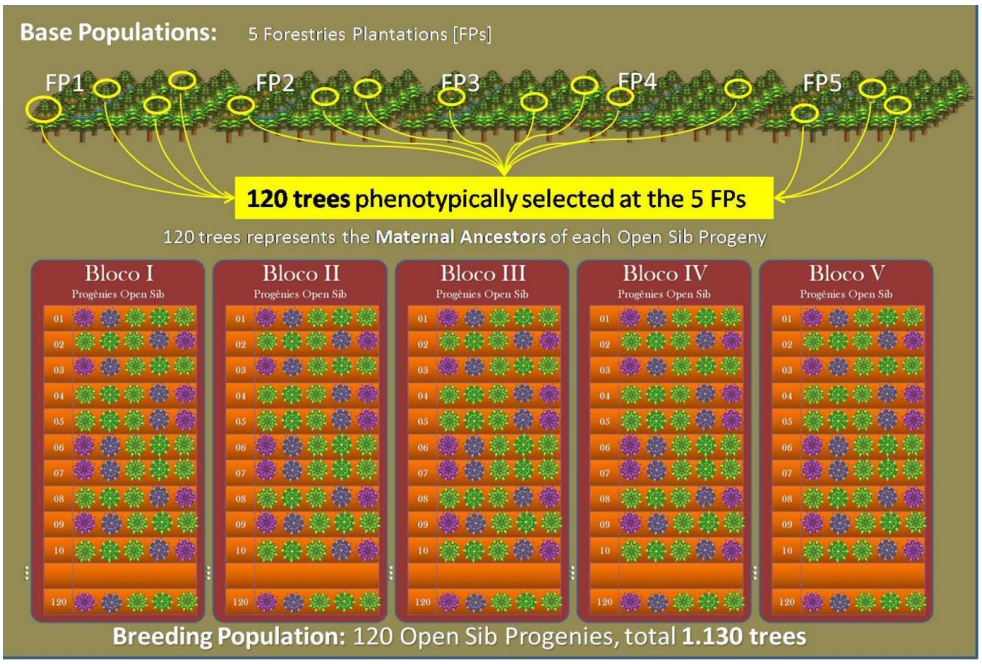
A figura 1 apresenta um esquema das populações de melhoramento existentes. A escolha da população de progênies de polinização aberta para iniciar o estudo de mapeamento de associação visando a caracterização de marcadores úteis à seleção assistida para qualidade da madeira, foram as seguintes:

i. A detecção de marcador útil à seleção assistida em forte desequilíbrio de ligação nessa população de polinização aberta, tem boa probabilidade de estar em desequilíbrio na outra de mesma origem: cruzamentos controlados.
ii. A probabilidade de encontrar alelos favoráveis para qualidade de madeira na população elite, que sofreu forte intensidade de seleção para outras características, é reduzida.
iii. Os indivíduos resultantes do cruzamento controlado estariam com dois anos de idade em 2009, impossibilitando fenotipagem para mapeamento.

Tabela 1. Número de progênies de polinização aberta – *Open-sib* (N_{os}), tamanho médio de progênie em número de indivíduos (N_{mp}), número total de indivíduos (N_{tp}) em cada Floresta Plantada (FP).

FP(cod.)	N_{os}	N_{mp}	N_{tp}	% size
AF	28	9,5	266	23,5%
B	18	9,27	167	14,8%
F	28	9,42	264	23,4%
L	5	9,2	46	4,1%
SV	41	9,44	387	34,2%
Total	120	9,42	1130	100%

Figura 1. Esquema da estrutura das populações de melhoramento disponibilizados para a presente pesquisa.



Fonte: Mercer, Juliane (2009), imagem criada e apresentada durante apresentação de relatório das atividades desta pesquisa à coordenação do projeto de Melhoramento *Pinus taeda* - FINEP.

2. ISOLAMENTO DE DNA GENÔMICO

O isolamento do DNA genômico de acículas congeladas seguiu o protocolo de Bittencourt *et al*. (2003)* que segue protocolo de Doyle & Doyle (1990), ajustando-o para um controle mais efetivo nos erros de manipulação de grande volume de amostras. A qualidade e a quantidade de DNA isolado foram averiguadas em gel de agarose 0,8% e marcador de tamanho o fago Hind III. Foram isoladas 1.800 amostras de DNA. O protocolo encontra-se em Anexo (Protocolo 1).

A nomenclatura adotada para discriminar um grande número de amostras foi planejada cautelosamente, adotando-se o seguinte padrão: numeral romano (primeira posição) representava o bloco (1 a 5 blocos); numeral arábico (segunda posição) – indicava a progênie de polinização aberta (1 a 120); letra em itálico minúscula "*i*" (terceira posição) designava o valor genético atribuído pelo ranqueamento do próprio programa de melhoramento.

Protocolo 1. Isolamento de DNA Genômico de *Pinus**

* *Dr.*[a] *Juliana Vitória Messias Bittencourt (2008 - Lamef UFPR).*

1.[a] etapa: Purificação

1. Macerar em cadinho com nitrogênio líquido 3 gramas de acícula congelada até a formação de um pó fino;

2. Transferir 35 mg do pó macerado para um *eppendorf* de 1,5 ml e acrescentar 600 uL de tampão de extração CTAB (2% CTAB; 1,4M NaCl; 20 mM EDTA; 100 mM Tris-HCl pH 8,0; 1% PVP), agitar por 5 minutos a mistura;

3. Incubar a 65ºC em banho-maria durante 30 minutos e a cada 10 minutos agitar as amostras para a homogeneização;

4. Deixar chegar à temperatura ambiente e acrescentar 800 uL de CIA (*clorofórmio/álcool isoamílico - 24:1*);

5. Agitar durante 5 minutos em agitador para *eppendorfes*;

6. Centrifugar por 5 minutos a 14.000 rpm à temperatura ambiente;

7. Transferir a fase orgânica (*fase superior*) para um microtubo novo (*1,5 mL*);

8. Adicionar um volume de CIA (*clorofórmio/álcool isoamílico - 24:1*);

9. Repetir novamente as etapas 6 e 7;

10. Coletar aproximadamente 300 uL do sobrenadante e transferir para um novo tubo;

2.[a] etapa: Precipitação

11. Adicionar 2,5 volumes de etanol (96%) gelado. Deixar descansar por 24 horas em *freezer* a -20^{0}C;

3.[a] etapa: Lavagem

12. Centrifugar a 14.000 rpm por 7 minutos;

13. Eliminar o sobrenadante recuperando o *pellet* e secar;

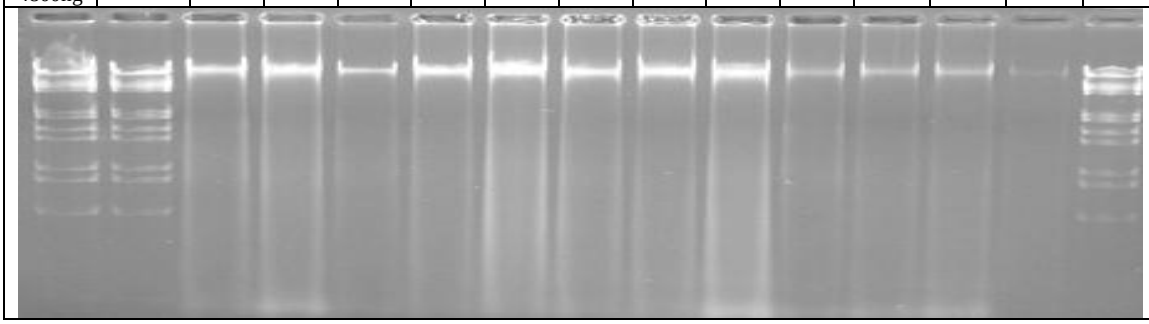
14. Ressuspender o *pellet* em 50 uL de TE 1x (Tris Hcl 1M pH 8,0; EDTA 0,5M);

15. Adicionar 1 ul de RNase (10mg/ml) e incubar a 37ºC por 30 minutos, armazenar a - 20ºC.

ANEXO II

Figura 1. Quantificação de DNA usando Hind III, marcador de banda conhecido.

M10	M5	1	2	3	4	5	6	7	8	9	10	11	12	M4
10ul	5ul	3450	2950	1650	2300	3600	2750	2950	4650	2200	2800	3400	3000	4ul
água 4800ng	2400ng	30	25.7	14.3	20	31.3	23.9	25.7	40.4	19.1	24.3	29.6	26.1	192ng



1	2	3	4	5	6	7	8	9	10	11	12
65 III ii	65 IV ii	65 V ii	66 I ii	66 II ii	66 III ii	66 IV ii	66 V ii	67 I ii	67 II ii	67 III ii	67 IV ii
67 V ii	68 I ii	68 II ii	68 III ii	68 IV ii	68 V ii	69 I ii	69 II ii	69 III ii	69 IV ii	69 V ii	70 I ii
70 II ii	70 III ii	70 IV ii	70 V ii	71 I ii	71 III ii	71 IV ii	71 V ii	72 I ii	72 III ii	72 IV ii	72 V ii
73 I ii	73 II ii	73 III ii	73 IV ii	73 V ii	74 I ii	74 II ii	74 III ii	74 IV ii	74 V ii	75 I ii	75 II ii
75 III ii	75 IV ii	75 V ii	76 I ii	76 II ii	76 III ii	76 IV ii	76 V ii	77 I ii	77 II ii	77 III ii	77 IV ii
77 V ii	78 I ii	78 II ii	78 III ii	78 IV ii	78 V ii	79 I ii	79 II ii	79 IV ii	79 V ii	80 I ii	80 II ii
80 III ii	80 IV ii	80 V ii	81 I ii	81 III ii	81 IV ii	81 V ii	82 I ii	82 II ii	82 IV ii	82 V ii	83 I ii
83 II ii	83 III ii	83 IV ii	83 V ii	84 I ii	84 II ii	84 III ii	84 IV ii	84 V ii	85 I ii	85 II ii	CONTROLE

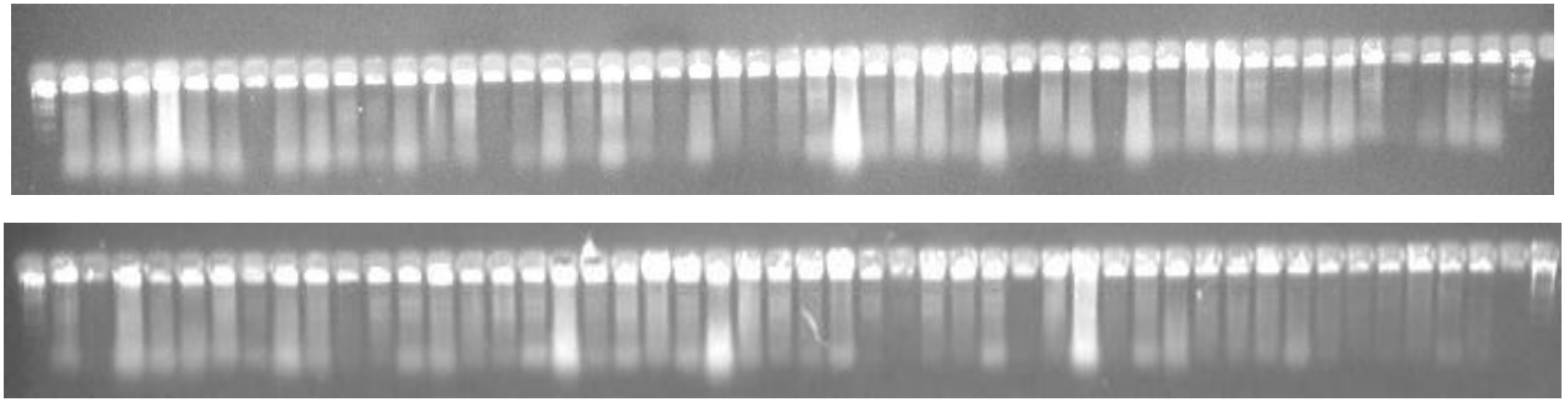


Figura 2. DNA de uma das 12 placas de 96 poços de indivíduos de polinização aberta da população selecionada de *Pinus taeda* L.

TABELA 1. *Primers* dos marcadores SSRs (microssatélies nucleares) polimórficos de *P. taeda.*

N.º *Primer Loci*	N.ºAcesso Gensbank	Dir	Seqüência	pb	Pb lit	Tm (ºC)	Fluoróforo	Motivo repetição	FONTE
1 PtTX 3026	AF 143971	F	AATACTTGGGAG GGATAC	350	335	58ºC	[HEX]l	$(ATC)_{11}…(ATC)_5…(ACC)_4(ATC)_6$	ELSIK *et al.* (2000)
		R	AATAGCCAGTTT TGTTTG		**344**				
3 PtTX 3002	AF 277846	F	TTGTTGTGCTCAT AATTACTAGTGT	200	194	58ºC	[6-FAM]	$(GAG)_6…(GAG)_4AA(GAG)_4$	ZHOU *et al.* (2003)
		R	CTCCTAAGCTTGC TCATGTG		**190**				
4 PtTX 3088	AF 277843	F	GGGCCTCCCTCC CCACAAAA	250	**244**	58ºC	[Cy3]	$(GAT)_5$	KUTIL & WILLIAMS (2001)
		R	TGGGATGGCTTG AGTTAAAGAACA		275				
6 PtTX 3091	AF 277848	F	GTGGCCACCTGC TTATT	220	**229**	58ºC	[HEX]	$(GTT)_{10}T_{13}(GGT)_{10}(CT)_5$	ZHOU *et al.* (2003)
		R	AACCCTTCCTATG ACTATGG		327				
7 PtTX 3098	AF 277847	F	TTTGCACTATGGC ATAAGTCCT	200	185	58ºC	[6-FAM]	$(GTT)_8$	KUTIL & WILLIAMS (2001)
		R	CCCTGTTTCTACC CTTGATGA		**187**				
8 PtTX 3118	AF277845	F	AACCATTTGCCC CTTCTT	200	216	58ºC	[Cy3]	$(CAT)_3. . .(CAT)_4. . .(CAT)_{11}$	KUTIL & WILLIAMS (2001)
		R	AAAAACAGTCTG CAATCAAATC		228				
9 PtTX 2037	AF 143959	F	GCCTTTAGATGA ATGAACCCA	180	**176**	58ºC	[HEX]	$(GTGA)_8(GT)_{14}$	ELSIK *et al.* (2000)
		R	TAAGCGGGATAT TATAGAGTTT		261				
17 PtTX 3025	AF 143970	F	CACGCTGTATAA TAACAATCTA	300	**266**	60ºC	[HEX]	$(CAA)_{10}$	ELSIK *et al.* (2000)
		R	TTCTATATTCGCT TTTAGTTTC		280				
18 RPtest5 (2c9f)	BV728798	F	ACAACAATAATA ACGGGGGC	200	**197**	60ºC	[Cy3]	$(AAC)_6$	CHAGNÉ *et al.* (2004)
		R	ACGCTTTAGATC CTCCTGCA						

CONTINUAÇÃO TABELA 1. *Primers* dos marcadores SSRs (microssatélies nucleares) polimórficos de *P. taeda.*

N.º *Primer Loci*	N.ºAcesso Gensbank	Dir	Seqüência	pb	Pb lit	Tm (ºC)	Fluoróforo	Motivo repetição	FONTE
19 2n11e	AA556153	F	AGGATTCCAACA GCATCACC	150	**147**	60ºC	[6-FAM]	$(TGC)_5$	ALLONA *et al.* (1998)
		R	CTGAACATGAAG CGCAGTGT						
20 RPtest8 (5c4g)	BV728799	F	GGTGCGAGATTG AAATTCGT	200	**196**	60ºC	[6-FAM]	$(CGC)_6$	CHAGNÉ *et al.* (2004)
		R	TTTGCAGTCTGTT GCCTTTG						
21 RPtest11 (6n7h)	BV728796	F	AGGATGCCTATG ATATGCGC	200	**213**	60ºC	[6-FAM]	$(ATC)_7$	CHAGNÉ *et al.* (2004)
		R	AACCATAACAAA AGCGGTCG						
22 8934M	AA739818	F	CAGAAATGGCGT CCAAATTC	150	**132**	60ºC	[6-FAM]	$[AGG]_5$	KINLAW, SC (1995) unpub
		R	ACCCCACTTATAT CCCCAGC						
23 9317M	AA740072	F	GTTCCCACTCAA GGGTTGAA	250	**259**	60ºC	[HEX] Verde	$(AGG)_3...(AGC)_5$	KINLAW, SC (1995) unpub
		R	ACATCATTTGTTG CCGCATA						
24 RPtest1 (1CA4D)	BV728795	F	GATCGTTATTCCT CCTGCCA	150	125	60ºC	[Cy3]	$(ATA)_7$	CHAGNÉ *et al.* (2004)
		R	TTCGATATCCTCC CTGCTTG						

ELSIK, C.G.; MINIHAN, V.T.; HALL, S.E.; SCARPA, A.M.; WILLIAMS, C.G. Low-copy microsatellite markers for Pinus taeda L. *Genome* 43: 550–555 (2000).
ZHOU, Y.; GWAZE, D.P.; REYES-VALDÉS, M.H.; BUI, T.; WILLIAMS, C.G. No clustering for linkage map based on low-copy and undermethylated microsatellites. *Genome* 46: 809–816 (2003).
KUTIL,B.L.; WILLIAMS,C.G. Triplet-repeat microsatellites shared among hard and soft pines. J. Hered. 92 (4), 327-332 (2001).
CHAGNÉ,D., CHAUMEIL,P., RAMBOER,A., ECHT,C.; PLOMION,C. *et al*. Cross-species transferability and mapping of genomic and cDNA SSRs in pines. JOURNAL Theor. Appl. Genet. 109 (6), 1204-1214 (2004). http://www.springerlink.com/content/v766ad0ryyk38yxa/fulltext.pdf
KINLAW,C.S. Loblolly pine cDNAs. 1995. Unpublished. Lab:USDA IFG. Dendrome Project of the Institute of Forest Genetics -Berkely.
ALLONA,I., QUINN,M., SHOOP,E., SWOPE,K., ST.CYR,S., CARLIS,J., RIEDL,J., RETZEL,E, CAMPBELL,M.M., SEDEROFF,R., WHETTEN, R.W. Analysis of xylem formation in pine by cDNA sequencing. **Proc. Natl. Acad. Sci.** U.S.A. 95 (16): 9693-9698. 1998. http://lifesciencedb.jp/bodymap-plant/download/library/fasta/PLN01847.fasta.gz

Dir: direção da sequência nucleotídica dos primers. / **F:** sentido a frente "foward". / **R**: sentido reverso "reverse". / **pb**: pares de bases nucleoticas do oligo sintetizado. / **pb lit**: pares de bases ou tamanho do motivo microssatélite encontrada na literatura. / Tm (0C): temperatura de anelamento utilizada no termociclador.

ANEXO III

ESTATÍSTICA *Ad Hoc* ΔK_M

A seguir, estão descritas as equações para estimar ΔK_m, que se referem a uma modificação realizada na equação da estatística *ad hoc* de EVANNO *et al*. (2005). Em suma, uma das únicas diferenças da estatística ΔK_m é usar a mediana no lugar da média para determinar as diferenças, em primeira e segunda grandeza, na representação das probabilidades *a posteriori* obtidas nas análises bayesianas para o parâmetro K dessa população. Esta modificação foi mais efetiva no sentido de tornar o pico modal do k_i, quando vários k_{ij}'s estão influenciando a análise.

O texto em adendo está escrito na língua inglesa, pois está sendo preparado para ser submetido como uma nota científica em uma revista ainda a escolher. No adendo, estão explicitadas as sequencias de atividades para realizar a análise estatística, desde a obtenção das probabilidades *a posteriori* para K, até a elaboração do gráfico para analisar o pico modal de ΔK_m.

ADENDO

ADAPTING THE *Ad Hoc* ESTATISTIC FOR FINDING THE TRUE K IN POPULATION GENETIC STUDIES - The Δk_m.

Abstract

The goal of the *ad hoc* statistic from Evanno *et al*. (2005) was to test the ability of the algorithm underlying the software STRUCTURE (Pritchard *et al*., 2000), to detect the number of clusters (K) in situations including more than two populations with different scenarios of migration patterns. While the program is increasingly used, it is unknown whether it can efficiently detect the real number of clusters in hierarchical systems, where

migration between populations is uneven. In the present note it was verified that a modification in the *ad hoc* statistic of the referred authors, by using the median values for the posterior distribution of K, instead of the mean values, allowed a better fit of the ΔK. Because of that, the ΔK described by the authors was namely in this note as ΔK_m. Description of the steps procedures considered in the modified *ad hoc* Statistic proposed by Evanno *et al*. (2005), and the graphical comparison for the both methods are discussed.

Methods

Population of Study and Multicluster Analysis

The reference family used in this study consisted of 120 open-sib progenies at 12 years old, installed in a complete block design with five repetitions. The total number of sample was 1,130 trees. Each progeny was originated in one of five forestry plantations (FPs) at Santa Catarina, a south state of Brazil, about 15 years ago. The seed cones that gave rise to the target population had been collected from 120 maternal ancestors, which had been phenotypically selected for wood volume at the five FPs.

The sample material was frozen needles tissue, kept in the −20ºC freezer. DNA was isolated using CTAB. A microsatellite multiplex system protocol was used for genotyping in MegaBace 1000 (GE Healthcare). The microsatellites *loci* used with their respective number of accession in Genbank, within parenthesis, were: ***PtTX3026*** (AF143971); ***PtTX3002*** (AF277846); ***PtTX3088*** (AF277843); ***PtTX3091*** (AF277848); ***PtTX3098*** (AF277847); ***PtTX3118*** (AF277845); ***PtTX2037*** (AF143959); ***PtTX3025*** (AF143970); ***RPtest5*** (BV728798); ***2n11e*** (AA556153); ***PRtest8*** (BV728799); ***PRtest11*** (BV728796); ***8934M*** (AA739818); ***9317M*** (AA740072) e **PRtest1** (BV728795).

The structure was characterized using bayesian inference in multicluster analysis with 15 microsatellite *loci* (SSR) in 1,130 samples, using the software STRUCTURE 2.2 Available at: <http://pritch.bsd.uchicago.edu/structure.html>). Most of parameters at the structure were set to their default values as advised in the user's manual of Structure 2.2 (Pritchard *et al*., 2004), also with the *admixture* model and the option of correlated allele frequencies between populations.

A pilot study was conduct, and a number of burn in 10,000 and its length of 50,000 each was enough. The range of possible *K*s tested was from 1 to 21, and for each *prior* K, 20

runs were carried out in order to quantify the amount of variation of the likelihood for each run.

Results and Discussion

The model choice criterion implemented in structure to detect the true *K* is an estimate of the posterior probability of the data for a given *K*, called the Pr(*X*|*K*) (Pritchard *et al* . 2000) and is called Ln P(D) in Evanno *et al*. (2005). The true number of populations (*K*) is often identified using the maximal value of Ln P(D) returned by structure.

It was observed (Table 1) as in Evanno *et al*. (2005) simulations, that in most cases, once the real *K* is reached, Ln P(D) stabilizes at larger Ks plateaus or continues increasing slightly. This phenomenon is also reported in the structure's manual (Pritchard *et al*., 2004), in which the variance between runs increases. The rational for this ΔK is to make salient the break in slope for the distribution of Ln P(D) at the true K.

First, it was estimated the mean likelihood Ln P(D) by the STRUCTURE consecutives analyses. From now on, Ln P(D) will be referring as *L*(*K*). Then, it was plotted the mean likelihood *L*(*K*) over 20 runs for each *K*. This mean was estimated using:

$$L(\bar{K}_k) = \sum_{j=1}^{i}(L(K_i))$$

Being, *i*: the prior *K* used in each run, $i = 1,\ldots,n$. (n=20);

j: the number of run repeated for the same prior *K*, $j = 1,\ldots,r$. (r=20).

Second, it was plotted the median difference between successive mean likelihood values of *K*, here will be called $L(\bar{K}_k)$. This difference corresponds to the rate of change of the likelihood function with respect to *K*, and is noted $L'(\bar{K}_k)$

$$L'(\bar{K}_k) = L(\bar{K}_k) - L(\bar{K}_{k-1})$$

In a third step it were plotted the absolute value of the difference between successive median values of $L'(\bar{K}_k)$. This corresponds to the second order rate of change of *L*(*K*), with respect to *K*.

$$|L''(\bar{K}_k)| = |[[\, L(\bar{K}_k) + L(\bar{K}_{k+1})] - 2.\, L(\bar{K}_k) + [\, L(\bar{K}_k) - L(\bar{K}_{k-1})]]|.$$

Finally, it was estimated ΔK as the mean of the absolute values of $L''(\hat{K}_k)$ averaged over 20 runs, divided by the standard deviation of the $L(K)$, that will be called here as $s[L(K)]$:

$$\Delta K = |L''(\hat{K}_k)| / s[L(K)]$$

It was divided $|L''(\hat{K}_k)|$ by $s[L(K)]$ because Evanno *et al*. (2005) detected a clear and general trend toward an increase of the variance of $L(K)$ between runs, as K increased. Also the authors concluded that a modal value for the distribution of ΔK to be located at the real K; and so, they used the height of this modal value as an indicator of the strength of the signal detected by STRUCTURE.

The result for the breeding population, found that the *ad hoc* quantity based on the second order rate of change of the likelihood function with respect to ΔK did show a clear peak at the $K=2$ and a second at K=4 (Figure 1), which is very similar with the patterns of the Stepping stone migration pattern simulated at Evanno *et al*. (2005).

By applying the modification in the *ad hoc* statistic, the modal values of ΔKm increased greatly, in comparison to the original ΔK, making much more salient the peak at the K=2, and a second at K=4 (Figure 2). Also, it was detected a third order peak for K=19, which was not detected by the original estimative of ΔK. At figure 3, the values of ΔK_m for K=2 were excluded from the graph plots for better visualizing the peaks of K=4 and K=19.

Table 1. Resumes the 200 runs resulted in STRUCTURE used for inferring the *ad hoc* value of Alpha.

Prior K	Mean LnP(D)	Minimum	Maximum	Mean deviation LnP(D)	Median deviation LnP(D)	Var. LnP(D)[1]
k 1	-18320,01	*-18320,30*	*-18319,90*	0,13	-18320,00	25,56
k 2	-16691,80	*-16729,50*	*-16666,10*	22,16	-16684,85	526,45
k 3	-16324,85	*-21462,40*	*-15520,00*	1812,63	-15865,70	1881,92
k 4	-14975,12	*-14991,50*	*-14951,50*	13,51	-14978,10	954,81
k 5	-14745,12	*-14897,30*	*-14654,00*	81,33	-14709,45	1187,24
k 6	-14392,52	*-14517,00*	*-14334,30*	66,22	-14363,15	1242,81
k 7	-14153,12	*-14289,30*	*-14092,60*	52,57	-14150,80	1428,53
k 8	-13914,55	*-13991,50*	*-13880,60*	34,76	-13903,20	1481,89
k 9	-13720,39	*-13878,30*	*-13648,00*	62,08	-13712,10	1564,21
k 10	-13598,41	*-13822,90*	*-13484,70*	107,02	-13583,60	1746,58
k 11	-13694,60	*-14795,40*	*-13372,70*	426,58	-13592,15	2309,48
k 12	-13598,57	*-14481,30*	*-13217,60*	374,23	-13477,20	2437,27
k 13	-13348,75	*-13618,10*	*-13204,00*	123,75	-13337,30	2189,96
k 14	-13324,10	*-13928,60*	*-13143,80*	238,26	-13232,80	2384,62
k 15	-13342,76	*-13669,00*	*-13079,20*	220,90	-13282,90	2648,82
k 16	-13231,30	*-13706,50*	*-13123,40*	172,08	-13179,15	2570,78
k 17	-13364,73	*-14191,50*	*-13107,10*	316,59	-13254,75	2969,09
k 18	-13537,15	*-14339,70*	*-13038,50*	451,31	-13296,15	3473,01

k 19	-13436,30	*-14436,50*	*-13086,80*	465,41	-13220,45	3396,41
k 20	-13607,79	*-16012,60*	*-13012,20*	923,77	13186,55	3791,30
k 21	-13211,64	*-13343,00*	*-13130,60*	67,70	-13202,50	3082,67

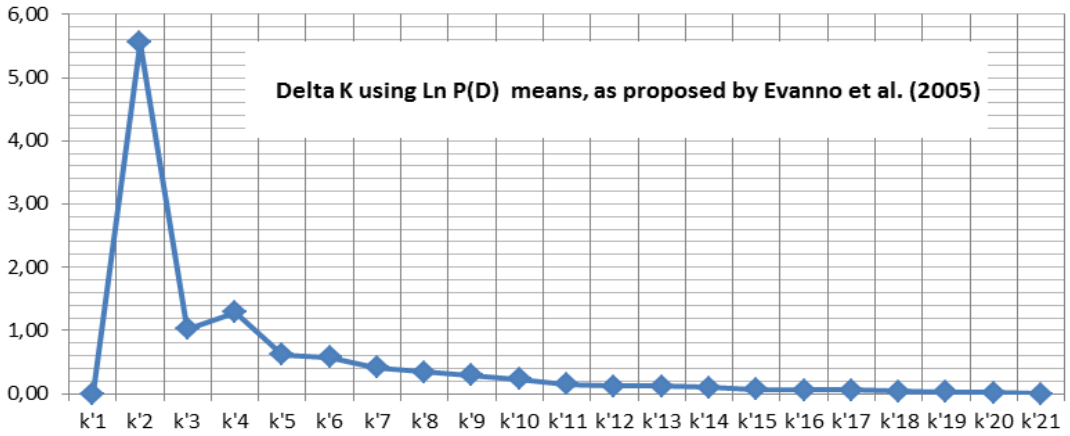


Figure 1. The graphic of the values of Δk plotted for the breeding population.

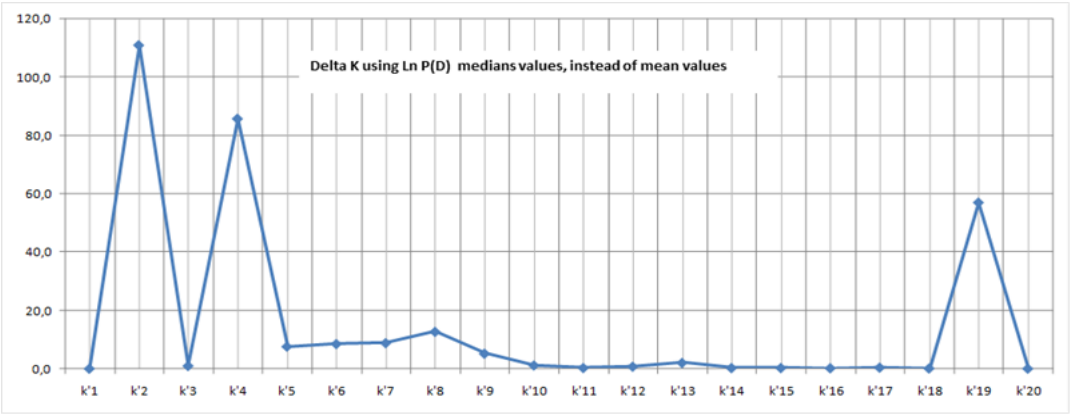


Figure 2. The graphic of the values of Δkm plotted for the breeding population.

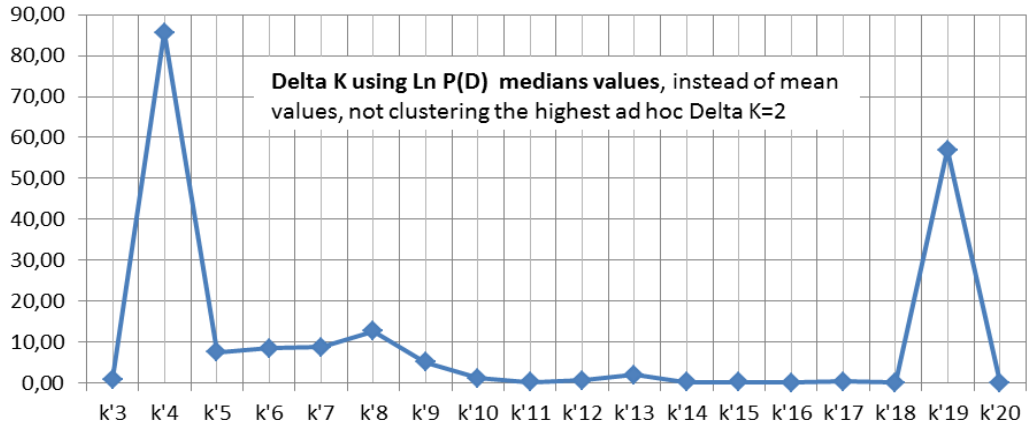


Figure 3. The graphic of the values of Δkm plotted by excluding the K=2.

Conclusion

It was concluded that the proposed modification for the *ad hoc* statistic of Evanno, by using the median values, may be more effective when the results for the posterior probabilities of the K, over the different runs of the STRUCTURE, are not well represented by the mean values as the central tendency of the data. Making the graphical plots for the ΔK_m much more salient than the original ΔK.

References

Evanno, G.; Regnaut, S.; Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*: **10**, 1365-1375. 2005.

Pritchard, J.K; Stephens, M.; Donnelly, P.J. Inference of population structure using *multilocus* genotype data. *Genetics*: **155**, 945-959. 2000.

Pritchard, J.K., Wen, W. **Documentation for Structure software version 2.** Department of Human Genetics. University of Chicago, Chicago, IL (available at <http://pritch.bsd.uchicago.edu>), 2004.

Pritchard, J.K.; Wen, X.; Falush, D. Documentation for structure software: Version 2.3. Software from http://pritch.bsd.uchicago.edu/structure.html. Feb., 2010.